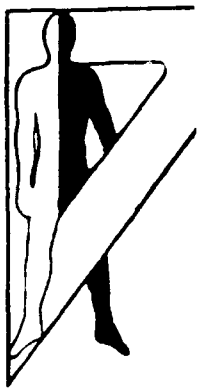


Technical Memorandum 14-91

AUTOMATIC RECOGNITION OF SPEECH IN STRESSFUL ENVIRONMENTS

Mark A. Clements
John H. Hansen
Kathleen E. Cummings
Sungjae Lim
Georgia Institute of Technology

August 1991
AMCMS Code 612716H700011

U.S. ARMY HUMAN ENGINEERING LABORATORY

Aberdeen Proving Ground, Maryland

Destroy this report when no longer needed.
Do not return it to the originator.

The findings of this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Use of trade names in this report does not constitute an official endorsement or approval of the use of such commercial products.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704-0188

1a. REPORT SECURITY CLASSIFICATION Unclassified	1b. RESTRICTIVE MARKINGS
2a. SECURITY CLASSIFICATION AUTHORITY	3. DISTRIBUTION/AVAILABILITY OF REPORT Approved for public release; distribution unlimited.
2b. DECLASSIFICATION/DOWNGRADING SCHEDULE	
4. PERFORMING ORGANIZATION REPORT NUMBER(S) Technical Memorandum 14-91	5. MONITORING ORGANIZATION REPORT NUMBER(S)
6a. NAME OF PERFORMING ORGANIZATION Human Engineering Laboratory 6b. OFFICE SYMBOL (If applicable) SLCHE	7a. NAME OF MONITORING ORGANIZATION
6c. ADDRESS (City, State, and ZIP Code) Aberdeen Proving Ground, MD 21005-5001	7b. ADDRESS (City, State, and ZIP Code)
8a. NAME OF FUNDING/SPONSORING ORGANIZATION 8b. OFFICE SYMBOL (If applicable)	9. PROCUREMENT INSTRUMENT IDENTIFICATION NUMBER
8c. ADDRESS (City, State, and ZIP Code)	10. SOURCE OF FUNDING NUMBERS PROGRAM ELEMENT NO. 6.27.16 / PROJECT NO. 1L162716AH70 / TASK NO. / WORK UNIT ACCESSION NO.

11. TITLE (Include Security Classification)
Automatic Recognition of Speech in Stressful Environments

12. PERSONAL AUTHOR(S) Clements, M.A., Hansen, J.H., Cummings, K.E., Lim, S.
Georgia Institute of Technology

13a. TYPE OF REPORT	13b. TIME COVERED	14. DATE OF REPORT (Year, Month, Day)	15. PAGE COUNT
Final	FROM _____ TO _____	1991, August	70

16. SUPPLEMENTARY NOTATION

17. COSATI CODES			18. SUBJECT TERMS (Continue on reverse if necessary and identify by block number)
FIELD	GROUP	SUB-GROUP	automatic recognition, robust ASR, enhancement, stress, noise removal
09	06		

19. ABSTRACT (Continue on reverse if necessary and identify by block number)

This document describes a body of research conducted for the purpose of improving the performance of automatic speech recognition (ASR) systems in noisy, stressful environments. To accomplish this, several avenues of basic research were investigated. First, novel techniques for noise reduction were developed. Second, three novel robust recognition systems were developed. One of these techniques relied on noise removal, another on linear least squares system identification, and the third on robust distance measures. Third, methods for characterizing stressed speech parametrically were developed. Among the descriptors used were various prosodic features, spectral features, and glottal excitation characteristics. Fourth, methods were developed that were shown to improve ASR performance in noisy or stressed conditions.

20. DISTRIBUTION/AVAILABILITY OF ABSTRACT ☐ UNCLASSIFIED/UNLIMITED ☒ SAME AS RPT. ☐ DTIC USERS	21. ABSTRACT SECURITY CLASSIFICATION Unclassified
22. NAME OF RESPONSIBLE INDIVIDUAL Technical Reports Office	22b. TELEPHONE (Include Area Code) (301) 278-4478 22c. OFFICE SYMBOL SLCHE-SS-TSB

DD Form 1473, JUN 86 *Previous editions are obsolete.* SECURITY CLASSIFICATION OF THIS PAGE

UNCLASSIFIED

AMCMS Code 612716H700011

Technical Memorandum 14-91

AUTOMATIC RECOGNITION OF SPEECH IN STRESSFUL ENVIRONMENTS

Mark A. Clements
John H. Hansen
Kathleen E. Cummings
Sungjae Lim
Georgia Institute of Technology

August 1991

APPROVED:

JOHN D. WEISZ
Director
Human Engineering Laboratory

U.S. ARMY HUMAN ENGINEERING LABORATORY

Aberdeen Proving Ground, Maryland

CONTENTS

FIGURES

TABLES

EXECUTIVE SUMMARY

This report describes a body of research directed toward improving automatic speech recognition (ASR) systems that operate in noisy and stressful environments. Several basic issues were addressed in this effort including (a) noise reduction, (b) robust recognition, (c) stress characterization, and (d) stress compensation for ASR. Also, an extensive stressed speech data base was compiled. The major results of the project are as follow:

1. Noise Reduction. A class of constrained iterative filtering algorithms was developed that displayed the following:

a. Better signal-to-noise ratios (SNRs) were observed than in other previously reported systems.

b. Improved numerical behavior was observed.

c. Improved automation was accomplished.

d. Objective quality measures were improved.

e. Consistent improvement across all manners of speech articulation (nasals, glides, vowels, etc.) was observed.

f. Subjective quality (informal tests) was significantly improved.

g. Significantly better ASR performance was observed when the speech was first processed by the system.

2. Recognition of Speech in Noise. Three systems were investigated:

a. Improved performance was observed when an ASR system was given as input to the speech processed as above.

b. A state space model of speech in noise and an ASR system designed specifically for this model (a continuous transition hidden Markov model [HMM]) were developed with significant performance improvement for short duration consonant identification resulting.

c. A projection-based distance function for ASR was investigated. Dramatic improvement was observed in even low SNRs, and at only a modest computational cost.

3. Stress Characterization. Descriptors for various stress styles were developed.

a. Prosodic features (e.g., intensity, pitch) were investigated and shown to be reliable indicators.

b. Spectral features (e.g., formants, spectral tilt) were investigated and were shown to display consistent features.

c. Glottal wave forms were extracted across a wide range of stress styles and were shown to be extremely consistent across different speakers, different vowels, and different utterances. For this task, a novel method of glottal extraction was developed.

4. Compensation. Modifications of ASR systems were made to improve performance in noisy and stressed conditions.

a. Noise-reduction techniques were shown to work on stressed speech.

b. Large improvements were observed in ASR performance with the assumption that the stress style could be identified--something that undoubtedly would be possible given the positive results of the characterization reported above.

5. Data Base. A large hand-segmented and labeled digital data base was compiled that included the Massachusetts Institute of Technology (MIT) Lincoln Labs multi-style data base, psychiatric patient recordings, recordings of speech in single and multiple tasking environments, and recordings of scripts spoken while in roller coaster and free fall conditions.

Although the reported project results are extremely useful in and of themselves, they have opened a number of doors suitable for future investigation.

AUTOMATIC RECOGNITION OF SPEECH IN STRESSFUL ENVIRONMENTS

INTRODUCTION

Although much effort has been expended in the task of automatic speech recognition (ASR), full acceptance of its implementation to solve practical problems has yet to be achieved. One of the main reasons for this current situation is the lack of robustness observed when high levels of interference (e.g., background noise, music, talking) are present or when the talker's tone of voice is changed. The terms "robust" and "robustness" refer to systems whose performance is not sensitive to small perturbations or adverse conditions. This latter condition is called *stress*. It is not difficult to imagine situations when a talker exhibits stress by talking louder, faster, or slower than usual while angry, during a heavy work load, and so forth. If an ASR device is to function successfully in an environment such as a helicopter cockpit, it will need to work in an environment with high noise levels whose characteristics may change substantially over time. It will also have to accept speech from a pilot who will be hearing the background noise, who will be under a high task load, who will sometimes be expressing fear and other emotions. This research project addressed many of the basic issues that must be better understood before adequate solutions would be possible. In particular, we have examined

1. Techniques for noise reduction. Here, a novel set of procedures that clean up noise-corrupted speech have been developed.

2. Robust recognition. Three novel techniques that aid in the automatic recognition of noise-corrupted speech have been developed.

3. Compilation of a stress data base. An existing data base has been substantially augmented to include speech from a large variety of speaking styles, stress conditions, and environmental situations.

4. Stress characterization. Many aspects of stressed speech have been examined including prosodics, spectral characteristics, and glottal wave forms. This last aspect required development of a novel extraction algorithm specifically tailored to stressed speech.

5. Compensation of stressed speech. The knowledge gained from the characterization aspect was used to improve ASR of stressed speech. The knowledge gained from the noise-treatment portions of the research was also combined into the algorithm for additional gains.

This document discusses the results of the research program, which has produced to date eight papers appearing in conferences in which the abstracts are reviewed (four photo-reduced mat pages each) (Hansen & Clements, 1987; Hansen & Clements, 1988; Hansen & Clements, 1989a; Hansen & Clements, 1989b; Carlson & Clements, 1990; Cummings & Clements, 1990; Cummings, Clements, & Hansen, 1989; Clements & Lim, 1987); one paper accepted for publication in a journal (Lim & Clements, 1990 [32 pages]); one paper to be submitted to a journal (Hansen & Clements, 1990 [14 pages]); and one Ph.D. thesis (Hansen, 1988 [430 pages]). Naturally, the discussion of the results in this report must be greatly compressed, and supporting documentation is heavily referenced.

ENHANCEMENT OF NOISY SPEECH

The success of a speech enhancement algorithm depends on the objectives made in deriving an approach. Assumptions made in this environment include (a) the noise distortion is additive; (b) only the degraded speech signal is available; and (c) the noise and speech signals are uncorrelated. The basis of the original unconstrained iterative enhancement approach is noncausal Wiener filters (Lim & Oppenheim, 1979). This approach tries to solve for the maximum likelihood estimate of a speech wave form in additive white gaussian noise with the requirement that the signal is the response from an all-pole process. Crucial to the success of this approach is the accuracy of the estimates of the all-pole parameters at each iteration. The algorithm is formulated by considering the case when all unknowns (all-pole speech parameters a, noise-free speech S_O) are random with *a priori* gaussian probability density functions. The basic procedure used is a maximum *a posteriori* (MAP) estimator that maximizes the probability density function of the unknown parameters given the noisy observations. After some simplification, it can be shown that the resulting equations for the joint MAP estimate of a and S_O become nonlinear, involving partial derivatives with respect to a. Lim and Oppenheim (1979) considered a suboptimal solution employing a sequential two-step approach based on MAP estimation of S_O followed by MAP estimation of a given $S_{O,i}$, in which $S_{O,i}$ is the result of the first estimation. This sequential estimation procedure is linear at each iteration and continues until some convergence criterion is satisfied. After further simplifying assumptions, it can be shown that the MAP estimation of S_O is equivalent to a minimum mean squared error (MMSE) estimate. In addition, as the observation window increases, the procedure for obtaining an MMSE estimate approaches a noncausal Wiener filter.

Although successful in a mathematical sense, this technique has received little application because of several factors. First, the scheme is iterative with sizable computational requirements. Second and most important is that although the original sequential MAP estimation technique was shown to increase the joint likelihood of the speech wave form and all-pole parameters, an heuristic convergence criterion had to be employed. This is a serious drawback if the approach is to be used in environments requiring automatic speech enhancement. After an extensive investigation (Hansen & Clements, 1985), this approach produced significant levels of enhancement for white gaussian noise in three to four iterations. Some interesting anomalies were noted that helped motivate development of the constrained approaches. First, as additional iterations were performed, individual formants of the speech decreased in bandwidth and shifted in location. Second, frame-to-frame pole jitter was observed across time. Both effects contributed to unnatural sounding speech. The goal, therefore, was to formulate a new set of enhancement algorithms that impose constraints on pole locations across time (inter-frame) and iterations (intra-frame). Spectral constraints are applied to the all-pole parameters $\hat{a}_i$, which ensure that (a) the all-pole speech model is stable, (b) it possesses speech-like characteristics (e.g., poles are not too close to the unit circle causing narrow bandwidths), and (c) the vocal tract characteristics do not vary wildly from frame to frame when speech is present. Because of the constraints imposed, improved estimates of $\hat{a}_{i+1}$ result. Given this new estimate, the second MAP estimation of S_O can be made. To increase numerical accuracy, reduce computational requirements, and eliminate inconsistencies in pole ordering across frames, the line spectral pair (LSP) transformation was used to implement most of the constraint requirements.

The LSP transformation can be viewed as an alternate representation of the LPC spectrum. The LSP coefficients are obtained from the linear predictive coder (LPC) prediction coefficients by combining the forward and backward predictor polynomials as follows:

$$P(z) = A(z) + B(z), \quad Q(z) = A(z) - B(z) \tag{1}$$

The vocal tract transfer function is given by $g/A(z)$, and M is the order of the LPC speech model. The resulting polynomials, P(z) and Q(z), are symmetrical and anti-symmetrical, respectively, with a root of P(z) at $z = +1$ and a root of Q(z) at $z = -1$. The remainder of the roots of P and Q lie on the unit circle. Since the roots occur in conjugate pairs, the original polynomial can be represented by M real numbers. The angles of the roots, $(\omega_i, i = 1,2,\ldots,M)$, are called the *line spectrum pairs*.

The LSPs possess several important properties that make them attractive for use in applying spectral constraints. One important characteristic is that if the vocal tract polynomial A(z) has all its roots inside the unit circle (i.e., a stable filter), the roots of P and Q will alternate around the unit circle (Crosmer, 1985). If two adjacent LSP frequencies are identical, it indicates that a root of A(z) lies on the unit circle.

In addition to their attractive representation of the LPC spectrum, the LSP coefficients offer the possibility of a more direct representation of perceptually important information. Specifically, there is a firm statistical relationship between the locations and bandwidths of the speech formants and the locations of the roots of P and Q, respectively. Since roots of the P polynomial correspond approximately to locations of formant center frequencies (when a formant is present), the P polynomials' LSP coefficients are termed *position coefficients*. It can be shown that the closer two LSP coefficients are together, the narrower the bandwidth of the corresponding pole of the vocal tract filter. Therefore, formants are indicated when two LSP coefficients are close together. When LSP coefficients are far apart, they indicate poles that contribute only to the overall spectral shape. Because of their relationship to the presence or absence of a formant by their nearness to a position coefficient, the coefficients of Q are termed *difference coefficients*. Given the LSP coefficients, the position coefficients are simply the odd index LSP coefficients, $(P_i = \omega_{2i-1}, i = 1,2,\ldots M/2)$. The difference coefficients are given as follows:

$$(|d_i| = \text{MIN}\ [|\omega_{2i+j} - \omega_{2i}|],\ i = 1,\ 2,\ldots,M/2) \qquad j = -1,1 \tag{2}$$

in which the sign of d_i is positive if ω_{2i} is closer to ω_{2i+j} and otherwise is negative. With this interpretation, a new enhancement technique based on Wiener filtering is now possible by imposing constraints on the LSP coefficients.

Figure 1 illustrates the framework for the constrained enhancement algorithms.

Results

Speech degraded by additive white gaussian noise was processed using various configurations of the new constrained enhancement algorithm. Energy thresholds for inter-frame constraints were obtained from frame energy histograms at each SNR. Excellent enhancement resulted for a wide range of threshold values. Intra-frame constraints were applied across two to three

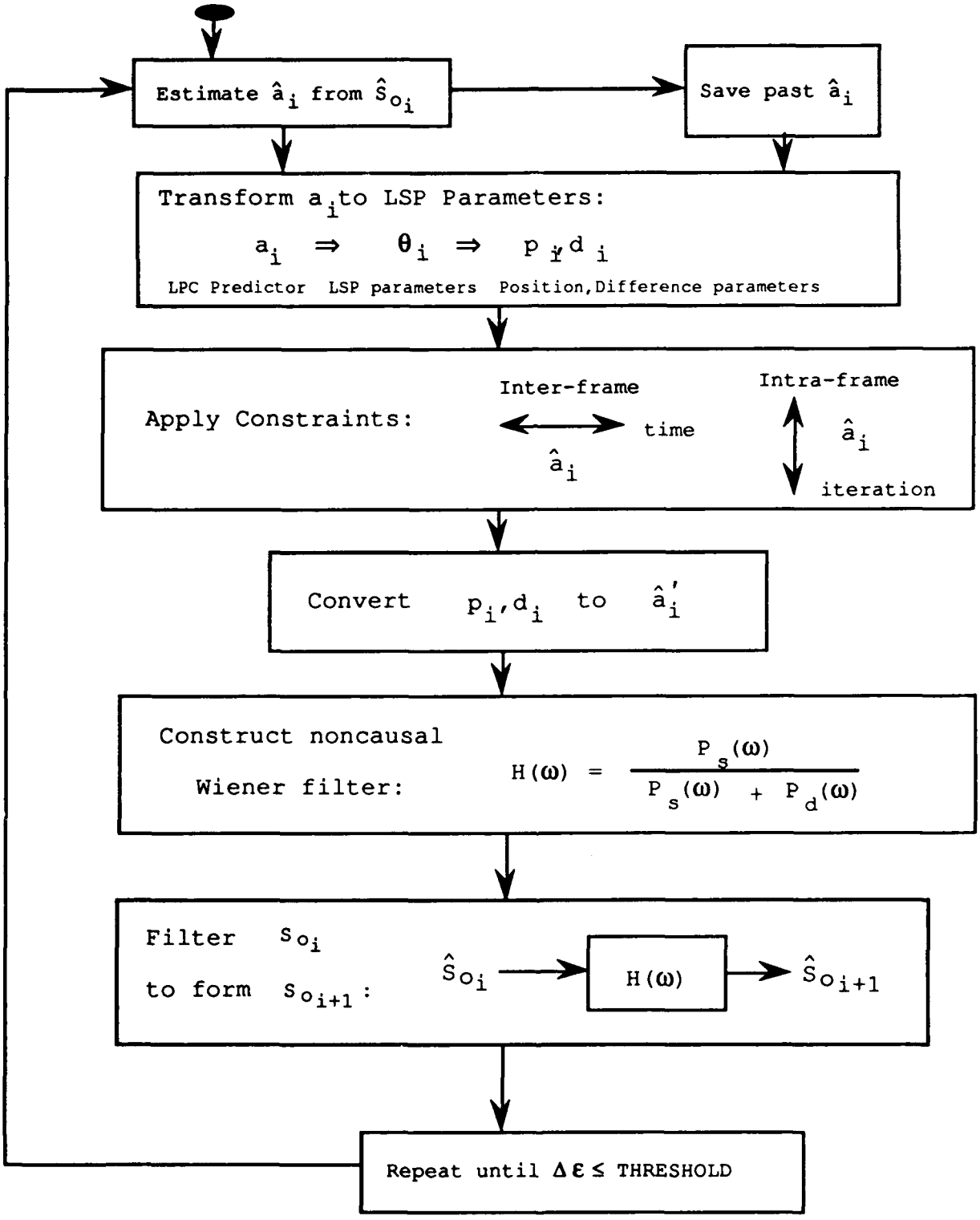


Figure 1. Framework for the constrained iterative enhancement algorithms.

iterations. Informal listening tests indicated noticeable quality improvement, although no intelligibility testing has been performed. However, extensive work has been performed in the area of objective speech quality measures (Quackenbush, Barnwell, & Clements, 1988). Good correlation has been shown to exist between subjective quality and objective measures. Therefore, objective measures including the Itakura-Saito likelihood ratio, log area ratio, and weighted spectral slope measure were used for evaluation.

Figure 2 illustrates a comparison of typical results for the various constraint approaches. The Itakura-Saito measure is plotted versus SNR for a white noise distortion. Plot a represents the original distorted speech. Plots b through e represent combinations of inter-frame constraints (both fixed and variable rate) and intra-frame constraints (applied to position coefficients-auto-correlation lags). All configurations examined showed significant improvement in Itakura-Saito measures. Threshold settings for the variable frame rate inter-frame constraint were somewhat sensitive to varying noise levels. However, the fixed frame approach by itself and with either auto-correlation or position intra-frame constraints, gave impressive results with little sensitivity to varying levels of SNR. To determine a limit of the level of enhancement, the original undistorted predictor coefficients, a, were used in the unconstrained algorithm. In essence, the two-step MAP estimation approach is now reduced to a single MAP estimate of S_O and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot f indicates this limit. Although only Itakura-Saito measures are shown, similar improvement was also observed for log area ratios and weighted spectral slope measures. Figure 3 compares the new approach to existing techniques. Plot b shows results from spectral subtraction as formulated by Boll (1979). An evaluation was performed for both half and full wave rectification, along with one to five frames of magnitude averaging, in which these points represent the best results. Plot c is from the unconstrained Wiener filtering technique. Plots d and e are typical values for the inter-frame constraint (fixed frame rate), and inter- plus intra-frame constraints (fixed frame and auto-correlation lags). Again, f indicates the limit for the Wiener filtering approaches.

Performance evaluation over sound classes was accomplished by hand-partitioning speech into segments. Entire sentences were processed, and objective measures from each class were computed. Table 1 summarizes this comparison between the unconstrained Lim-Oppenheim technique to that of the inter and intra-frame constraint approach. Measures for the theoretical limit using undistorted LPC predictor coefficients, a, are also indicated. Improvement is indicated for all types of speech. In addition, the constrained approach produced superior objective measures of quality across all speech classes at the same iteration. These results clearly indicate improvement over the unconstrained approach as well as spectral subtraction for additive white gaussian noise.

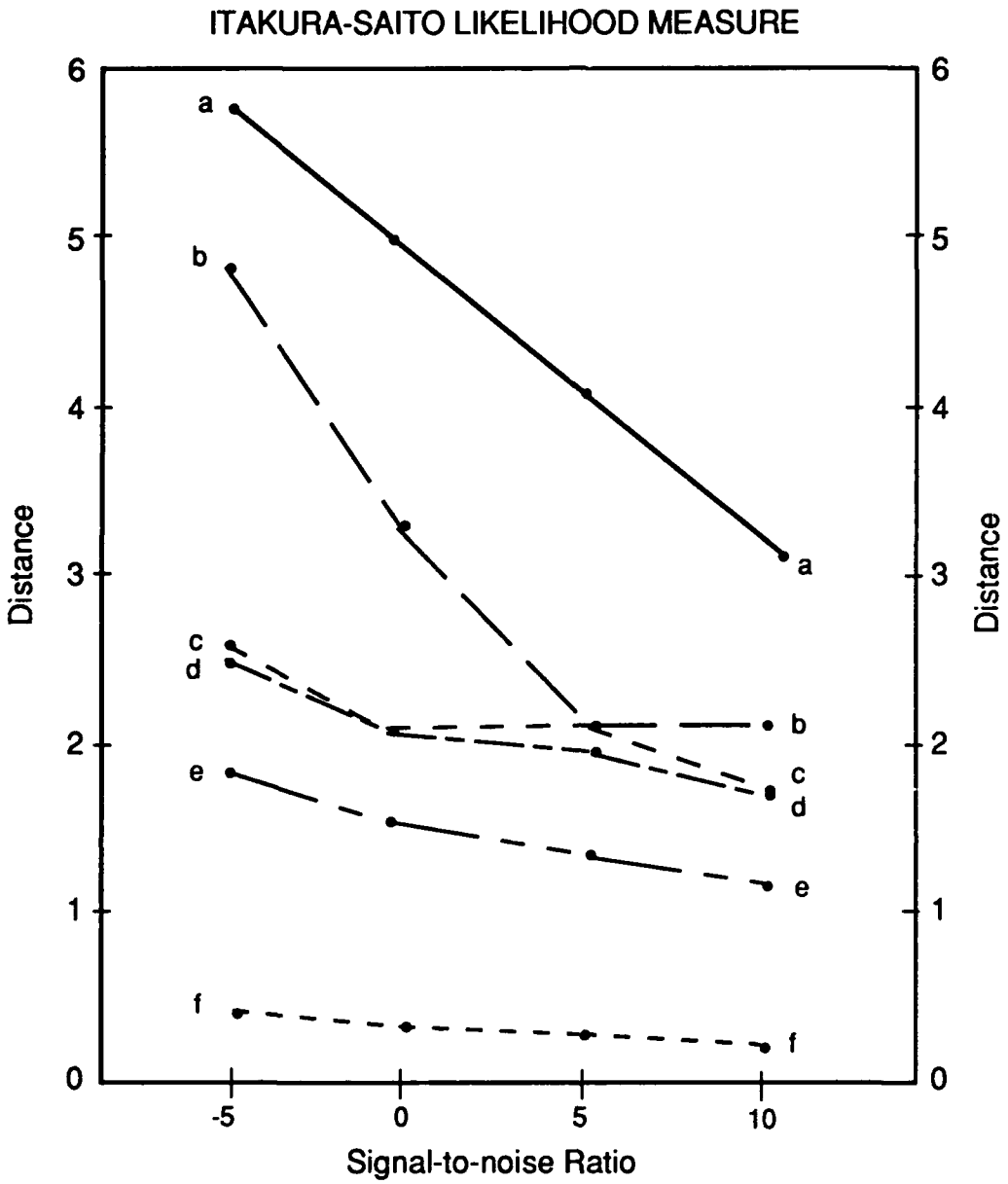


Figure 2. Comparison of constraint algorithms over SNR.

a. Original distorted speech

b. Inter-frame constraint: variable frame (VF-LSP:T)

c. Intra-frame constraint: fixed frame (FF-LSP:T)

d. Inter- and intra-frame constraints: fixed frame, position (FF-LSP:T,LSP:I)

e. Inter- and intra-frame constraints: fixed frame, auto-correlation (FF-LSP:T,Auto:I)

f. Theoretical limit: 0 using undistorted LPC coefficients a.

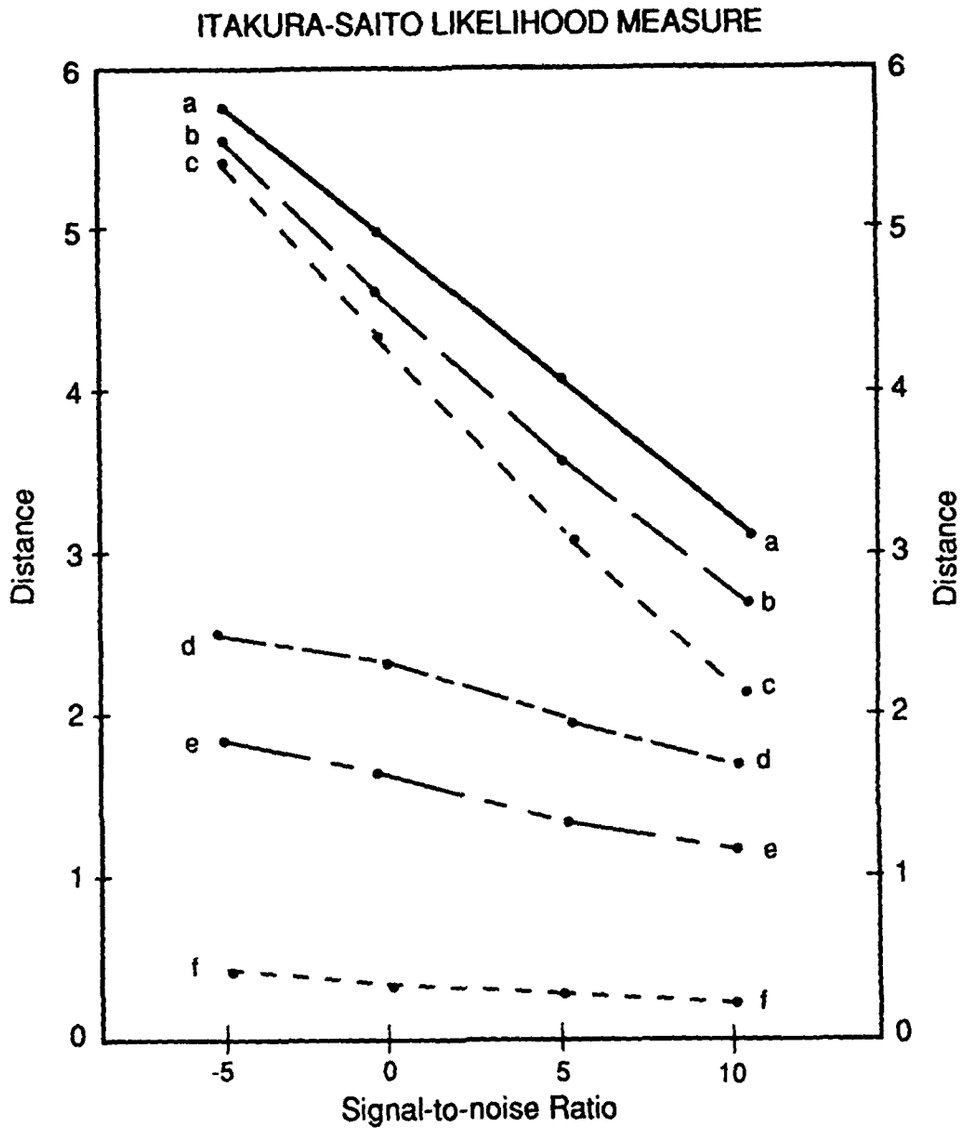


Figure 3. Comparison of enhancement algorithms over SNR.

a. Original distorted speech

b. Boll: spectral subtraction using magnitude averaging

c. Lim-Oppenheim: unconstrained Wiener filtering

d. Hansen-Clements: employing inter-frame constraints (FF-LSP:T)

e. Hansen-Clements: employing inter- and intra-frame constraints (FF-LSP:T,Auto:I)

f. Theoretical limit: using undistorted LPC coefficients a.

Table 1

Comparison of Unconstrained (Lim-Oppenheim) and Inter- and Intra-frame Constrained (Hansen-Clements) Algorithms Over Sound Types for White Gaussian Noise (SNR = +5 dB)

Sound type	Itakura-Saito likelihood measure			
	Original	Lim-Oppenheim	Hansen-Clements	True LPC
Silence	1.634	1.649	0.842	0.319
Vowel	4.020	3.299	1.651	0.582
Nasal	19.814	17.656	3.968	0.324
Stop	7.261	3.979	1.099	0.435
Fricative	3.739	3.509	1.766	0.649
Glide	1.525	1.442	1.131	0.705
Liquid	9.597	4.545	0.998	0.303
Affricate	3.924	2.702	2.229	0.323
Voiced and unvoiced	5.838	4.293	1.761	0.519
Total	4.022	3.151	1.364	0.433

As mentioned, the iterative enhancement algorithms must be suspended at some iteration. To determine a terminating iteration, a criterion must be must be selected to evaluate levels of improvement as the iterative scheme progresses. The criterion chosen is based on objective speech quality measures. Such measures are formed by a weighted comparison of actual and resulting estimated LPC predictor coefficients found during enhancement. The obvious problem with such a criterion is that except for simulation, the actual speech is unknown during the procedure. If, however, simulation were to show a consistent value for the best iteration in terms of this criterion, a convenient stopping condition would exist. Previous results based on objective quality measures indicate the unconstrained approach to produce maximum objective quality at different iterations for different classes of speech. Table 2 illustrates this behavior over the indicated sound classes. As this table shows, maximum overall speech quality is obtained at the third iteration, with considerable variation across sound types. For example, glides required two iterations, with nasals, liquids, and affricates requiring between five and six. Therefore, depending on sound class concentration, the optimal iteration (in terms of minimum distance) would vary considerably. This result indicates the inability to determine in advance a terminating iteration for the unconstrained approach since it is highly dependent on sound class and to a lesser degree on SNR.

The new constrained enhancement algorithms appear to solve this problem of sound class dependency. Table 3 presents results from an equivalent evaluation for one of the constrained enhancement algorithms (FF-LSP:T,Auto:1). A comparison between Tables 2 and 3 shows that the constrained approach produces superior quality measures across all speech classes at the same iteration. This improvement surpasses even combined individual maximum quality measures across the unconstrained approach. Thus, the constrained enhancement algorithm does more than simply impose a constraint to adjust the rate of improvement: the constrained approaches consistently result in superior objective speech quality at the same iteration over all sound

classes, independent of SNR. Table 4 summarizes optimum terminating points in terms of objective quality for the enhancement algorithms. Techniques employing only inter-frame constraints consistently resulted (93% occurrence) in maximum quality at the third iteration. Techniques employing inter- and intra-frame constraints had a 97% occurrence of maximum quality at the seventh iteration. In addition, adjacent iterations differ only slightly in objective quality for the constrained techniques. This contrasts sharply with the large variations in adjacent iterations for the unconstrained technique. Therefore, if the iterative scheme were allowed to continue or halted one iteration before optimal, only minor differences in speech quality would result. The results consistently suggested that the constrained enhancement algorithms reach a maximum level of speech quality at the same iteration, independent of SNR and sound class concentrations.

The unconstrained Wiener filtering-all-pole modeling approach was previously generalized for colored aircraft noise (Hansen & Clements, 1990). In that study, an extensive investigation was performed using various spectral estimation techniques (MEM, MLM, Burg, Bartlett, Pisarenko, Periodogram) for securing estimates of colored background noise, along with varying SNR (-20 dB to +20 dB). Results indicated that Bartlett's method produced spectral estimates that resulted in highest quality improvement for this particular distortion.

Table 2

Lim-Oppenheim Unconstrained Speech Enhancement for AWGN, STR=+5dB
(Optimum perceived quality for a particular speech class is indicated by a ♣.)

Sound type	Itakura-Saito likelihood measure (across iterations)							
	Original	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7
Silence	1.634	1.615	♣1.608	1.649	1.933	3.756	20.360	49.884
Vowel	4.020	3.721	3.445	♣3.299	3.720	8.319	121.82	---
Nasal	19.814	19.154	18.416	17.656	17.009	16.593	♣15.192	15.697
Stop	7.261	6.114	4.926	3.979	♣3.822	6.889	25.515	29.694
Fricative	3.739	3.637	3.532	♣3.509	3.902	7.658	47.829	94.106
Glide	1.525	1.414	♣1.333	1.442	2.231	4.300	8.391	15.561
Liquid	9.597	8.241	6.546	4.545	2.606	♣1.676	6.381	30.001
Affricate	3.924	3.609	3.213	2.702	2.091	♣1.552	2.911	2.975
Voiced and unvoiced	5.838	5.321	4.767	4.293	♣4.289	7.346	61.865	---
Total	4.022	3.720	3.402	♣3.151	3.271	5.795	43.457	---

Table 3

Hansen-Clements Inter- and Intra-frame Constrained Speech Enhancement for AWGN, SR=+5dB (Optimum perceived quality for a particular speech class is indicated by a ♣.)

Sound type	Itakura-Saito likelihood measure (across iterations)								
	Original	No. 1	No. 2	No. 3	No. 4	No. 5	No. 6	No. 7	No. 8
Silence	1.634	1.551	1.351	1.155	1.036	0.979	0.929	♣0.884	0.901
Vowel	4.020	3.319	2.865	2.394	1.863	1.677	1.571	♣1.565	1.828
Nasal	19.814	16.490	12.397	10.523	8.682	6.840	4.929	♣3.789	5.548
Stop	7.261	6.246	4.840	3.492	2.668	1.812	1.383	♣1.129	1.435
Fricative	3.739	3.432	3.027	2.612	2.245	1.948	1.729	♣1.615	1.844
Glide	1.525	1.389	1.275	1.232	1.219	1.189	1.161	♣1.153	1.217
Liquid	9.597	6.481	3.382	2.243	1.612	1.209	0.943	♣0.926	1.211
Affricate	3.924	3.722	3.447	3.117	2.806	2.598	2.472	♣2.368	3.966
Voiced and unvoiced	5.838	4.642	3.658	3.006	2.501	2.131	1.865	♣1.740	1.953
Total	4.022	3.026	2.441	2.069	1.801	1.611	1.457	♣1.381	1.498

Table 4

Summary of Optimal Terminating Iteration Across SNR for AWGN

Constrained enhancement algorithm	Additive white gaussian noise SNR: -5 dB		-0 dB		+5 dB		+10 dB		OVERALL	
	Optimal iteration using Itakura-Saito likelihood measure									
	Iter.	Freq. (%)	Iter.	Freq. (%)	Iter.	Freq. (%)	Iter.	Freq. (%)	Iter.	Freq. (%)
FF-LSP:T	3	100	3	87	3	87	3	100	3	93
			4	13	4	13			4	7
VF-LSP:T	3	90	3	85	3	94	3	100	3	94
	4	10	4	15	4	6			4	6
FF-LSP:T, Auto:I	7	100	7	100	7	100	7	88	7	97
							6	12	6	3
FF-LSP:T, LSP:I	4	100	4	100	4	100	4	100	4	100
VF-LSP:T, LSP:I	4	100	4	100	4	100	4	100	4	100

Noise recorded from a Lockheed C-130 aircraft interior was used to degrade noise free utterances. For these simulations, two Bartlett spectral estimates from the original noise wave form (to avoid complications in silence detection) were used across each sentence. The noise was both colored and non-stationary, so increasing the number of spectral estimates across the utterance should improve enhancement performance. An analysis.as performed for an inter-frame (FF-LSP:T) and a combined inter- and intra-frame (FF-LSP:T,Auto:I) approach. Informal listening tests indicated noticeable quality improvement. Figure 4 illustrates results from this study. All configurations examined-showed significant improvement in Itakura-Saito measures. Plot a shows Itakura-Saito measures for the original distorted speech. Plot b is from the unconstrained Wiener filtering technique. Plots c and d are typical values for the inter-frame constraint (FF-LSP:T), and inter- plus intra-frame constraint (FF-LSP:T, Auto:I) approaches. To determine limits of the level of enhancement, the original undistorted predictor coefficients were used in the unconstrained algorithm. In essence, the two-step MAP estimation approach is now reduced to a single MAP estimate of S_O, and therefore represents the theoretical limit for enhancement using Wiener filtering. Plot e indicates this limit. Although only Itakura-Saito measures are shown, similar improvement was observed for log area ratio and weighted spectral slope distance measures. As this figure indicates, significant levels of enhancement result for the constrained enhancement algorithms.

These results show that the constraint algorithms outperform the unconstrained approach for a colored distortion. However, it is possible that the constrained techniques are improving only particular speech classes which may have high concentrations in the test utterances. Therefore, a performance evaluation over sound classes was performed by hand-partitioning speech into segments, enhancing entire sentences, and computing objective measures from each class. Table 5 summarizes this comparison between the unconstrained technique to that of the inter- and intra-frame constraint approach (FF-LSP:T,Auto:I). Measures for the theoretical limit using undistorted LPC coefficients are also indicated. It should be noted that voiced plus unvoiced measures give a better indication of quality improvement because of the time-varying nature of the interfering background noise. Improvement is indicated for all types of speech. This shows that the constrained techniques are enhancing all aspects of the speech signal.

RECOGNITION OF SPEECH IN NOISE

Three approaches were investigated for recognition of speech in noise. In the first, the enhancement algorithms of Section I were directly applied to the noisy speech, and then input to an ASR module. In the second, a signal detection approach based on Kalman filter and a pseudo-continuous hidden Markov model (HMM) were used. In the third, a front end and distance measures were used, which in themselves do not enhance speech, but nevertheless improve performance because of the statistical properties of speech in noise.

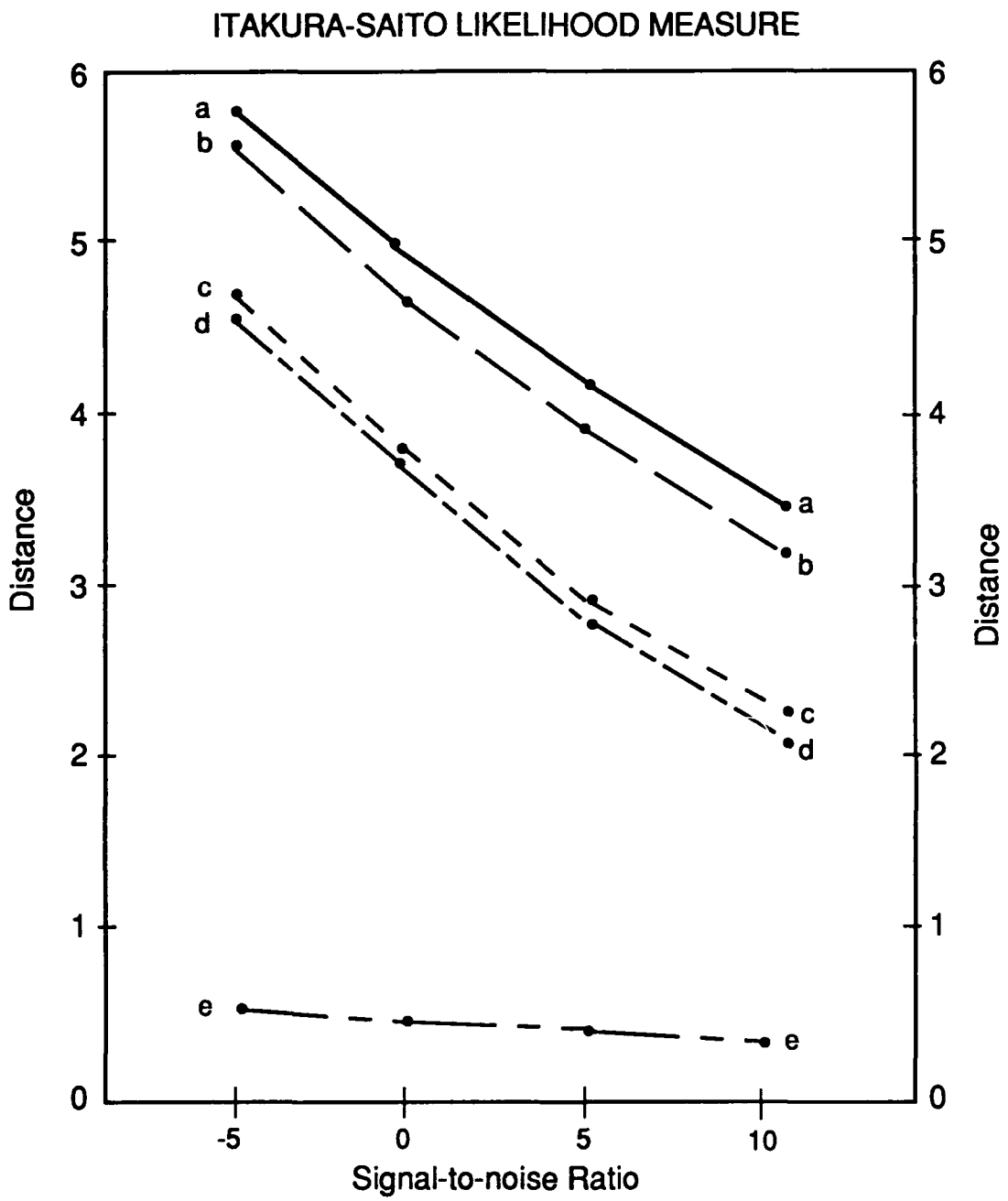


Figure 4. Comparison of inter- and intra-frame constrained enhancement algorithms for colored aircraft noise over SNR.

a. Original distorted speech

b. Generalized unconstrained Wiener filtering

c. Hansen-Clements: employing inter-frame constraints (FF-LSP:T)

d. Hansen-Clements: employing inter- and intra-frame constraints (FF-LSP:T,Auto:I)

e. Theoretical limit: using undistorted LPC coefficients $\vec{a}$.

Table 5

Comparison of Unconstrained (Lim-Oppenheim) and Inter- and Intra-Frame Constrained (Hansen-Clements) Algorithms Over Sound Types for Slowly Varying Colored Noise (SR=+5dB)

Sound type	Itakura-Saito likelihood measure			
	Original	Lim-Oppenheim	Hansen-Clements	True LPC
Silence	6.63	6.33	4.32	2.03
Vowel	3.23	2.54	1.44	0.53
Nasal	4.03	3.26	2.13	0.45
Stop	1.58	1.29	0.66	0.61
Fricative	1.37	1.09	0.85	0.65
Glide	1.14	1.04	0.52	0.51
Liquid	1.22	0.55	0.22	0.18
Affricate	0.90	0.51	0.33	0.16
Voiced and unvoiced	2.27	1.76	1.08	0.52
Total	4.15	3.86	2.74	1.17

Enhancement Followed by Recognition

The utility of the previously described enhancement procedures for helping an unmodified .SR module was tested. A fairly standard, isolated word, discrete observation HMM recognition system was used for evaluation. This system was LPC-based and had no embellishments. In all experiments, a five-state, left-to-right model was used. The system dictionary consisted of 20 highly confusable words used by Texas Instruments and MIT Lincoln Labs to evaluate recognition systems. Subsets include {go, oh, no, hello} and {six, fix}. Twelve examples of each word were used, six for training, six for recognition (i.e., all tests fully open). A vector quantizer was used to generate a 64-state codebook using 2 minutes of noise-free training data. The 20 models employed by the HMM recognizer were trained using the forward-backward algorithm. Table 6 presents results from five scenarios using a noise-free codebook and noise-free trained system. Spectral subtraction preprocessing employed three frames of magnitude averaging. The unconstrained Lim-Oppenheim approach was terminated at the third iteration. The constrained Hansen-Clements (FF-LSP:T,Auto:I) was terminated at the seventh. As these results indicate, recognition was reduced to chance for noisy, spectral subtraction, and Lim-Oppenheim (-5,0,5 dB) speech. The constrained approach resulted in improved recognition across all SNRs considered, which is quite remarkable in light of the severe levels of noise and the difficulty of dictionary employed. However, reliable recognition in such a hostile environment may require more than merely extending existing techniques. As a final comparison, three tests were performed using noisy and enhanced speech (SNR=+10 dB). For the noisy case, speech was coded using a noisy codebook, and recognition performed using a noisy trained HMM recognizer. Similar tests were performed for two enhancement techniques (i.e., enhanced words coded using enhanced codebook, and tested using enhanced speech trained HMM recognizer). Forty percent of the errors recognized were caused by misclassification of leading consonants (especially fricatives).

Table 6

Recognition of Performance Using Enhancement Pre-processing in AWGN
(†SNR=+10dB)

Condition (noise-free training)	RECOGNITION RESULTS Signal-to-noise ratio					
	Original (%)	-5 dB (%)	-0 dB (%)	+5 dB (%)	+10 dB (%)	+20 dB (%)	+30 dB (%)
Noise free	88						
Noisy		5	5	6.7	5	8	49
Spectral subtraction		5.8	7.1	5	5.4	20	55
Lim-Oppenheim		5.4	5.8	7.5	12.5	41	64
Hansen-Clements		15	14	19.5	34.5	59	83

Train and recognize in the same environment			
Noise free	Noisy †	Hansen-Clements †	Lim-Oppenheim †
88%	90%	77%	23%

The constrained algorithms have shown improvement as a preprocessor for speech recognition, although their ability to improve performance to an acceptable level in SNRs as low as those considered is questionable. Although the enhancement procedures improved LPC parameter estimation substantially, LPC-based strategies may simply be inappropriate for SNRs of roughly 0 dB. Further work in this SNR range will require as a minimum, different front end processing.

Recognition of Speech in Noise

In this section, a set of techniques based on linear least squares estimation theory is described. The basic idea is that the speech signal (in quiet or in noise) carries information about the system that produced it. By using optimal estimation and system identification, presumably one could improve ASR.

The motivation for the current study came from observations that although general performance of a recognizer may not depend highly on the exact placement of frames, the detailed error patterns often do. The methods explored try to eliminate the apparent framing artifacts by essentially extracting a set of parameters for every sample of the digital speech. The recognition algorithm can then be considered a close approximation of a continuous transition HMM. This approach would not be feasible were it not for the efficient algorithms formulated for this specific problem.

In this section, the aspects of HMMs, which are conducive to this strategy and issues involved in training, and recognition are discussed. Three parameter extraction methods, one of which relies on a novel use of

Kalman filtering, and others (two involving more classical procedures) are described. Experimental results are described.

The Hidden Markov Model

Definitions

Consider a discrete state discrete transition HMM for each pattern to be recognized. Assume the observations are drawn from a finite alphabet of size M, and a new observation is made for every sample of the digital speech. This would imply some form of vector quantizer continuously outputting a code word sequence. Although the form and implementation of this process are described in detail in the Recognition System section for all systems considered, enough memory existed in the analysis to produce long sequences of the same code word in a segment of an utterance. The importance of this result will become apparent below. Denote the number of states in a model by *n*.

π_i = probability the model starts in state i,

$\Pi^T = \pi_1, \pi_2, \ldots, \pi_n$

$\mathbf{A}$ = transition probability matrix, in which

a_{ij} = probability of transition from state i to state j in one trial; i, j = 1, 2,...,n.

$\mathbf{B}$ = observation probability matrix in which

b_{jk} = probability of observing code word k given state j.

$\mathbf{O}(t)$ = code word observed at time t, $1 \leq t \leq F$

$\mathbf{R}(t)$ = observation matrix, consisting of

$$\mathbf{R}(t) = \text{diag}[b_1(\mathbf{O}(t)), \ldots, b_n(\mathbf{O}(t))]$$

For a given model M, and observations $\mathbf{O}(1), \mathbf{O}(2), \ldots, \mathbf{O}(F)$, define

$\underline{\alpha}^T(t) = [\alpha_1(t), \ldots, \alpha_n(t)]$

$\alpha_i(t) = \text{prob}(\mathbf{O}[1], \ldots, \mathbf{O}[t];$ state i at t)

$\underline{\beta}^T(t) = [\beta_i(t), \ldots, \beta_n(t)]$

$\beta_i(t) = \text{prob}(\mathbf{O}[t+1], \ldots, \mathbf{O}[F];$ state i at t)

The probability that the sequence from the model is observed is

$$\Pr[\mathbf{O}(1), \ldots, \mathbf{O}(F)] = \sum_{i=1}^{n} \alpha_i(t)\beta_i(t) \qquad (3)$$

$\underline{\alpha}(t)$, $\underline{\beta}(t)$, and Equation (1) can be rewritten in matrix form so that

$$\Pr[O(1),\ldots,O(F)] = \Pi^T \mathbf{R}(1)\mathbf{A}\mathbf{R}(2)\mathbf{A}\ldots\mathbf{A}\mathbf{R}(F)\underline{\beta}(F) \quad (4)$$

$$\underline{\alpha}^T(t) = \Pi^T \mathbf{R}(1)\mathbf{A}\mathbf{R}(2)\ldots\mathbf{A}\mathbf{R}(t) \quad (5)$$

$$\underline{\beta}^T(t) = \mathbf{A}\mathbf{R}(t+1)\mathbf{A}\mathbf{R}(t+2)\ldots\mathbf{A}\mathbf{R}(F)\underline{\beta}(f) \quad (6)$$

If the model is constrained to be left to right, A will be upper triangular. If the model demands the system to start in state 1 and end in state n, then

$$\Pi^T = [1,0,\ldots,0] \quad (7)$$

$$\underline{\beta}^T(F) = [0,\ldots,0,1]$$

Recognition

For a given model, one needs to compute the probability of the observations. This can be accomplished by evaluating the above equations. In this system, F is normally such a large number that the direct evaluations would require an inordinate amount of computation. To reduce this computational burden, the fact that usually long runs of the same code words occur is employed to make the above equations a sequence of the same matrix multiplications. The constraint that the model be left to right that makes **A** upper triangular. Assume that the code words at time t + 1 through t + m are the same. The partial product for the period of time is

$$[\mathbf{A}\mathbf{R}(t+1)\mathbf{A}\mathbf{R}(t+2)\ldots\mathbf{A}\mathbf{R}(t+m)],$$

and is equal to

$$[\mathbf{A}\mathbf{R}(t+m)]^m$$

Since the matrix **A** is upper triangular and **R**(t + m) is diagonal, the product, $[\mathbf{A}\mathbf{R}(t+m)]^m$ is an upper triangular matrix. The upper triangular matrix has the property that it can be diagonalized if the diagonal elements are distinct. In this case, if the diagonal elements of **AR**(t + m) are assumed to be distinct, it can be diagonalized in such a form that

$$\mathbf{A}\mathbf{R}(t+m) = \mathbf{P}\mathbf{D}\mathbf{P}^{-1} \quad (8)$$

in which D is diagonal with its elements the same as the diagonal elements of AR(t + m); P is upper triangular with its diagonal elements equal to 1. Therefore,

$$[\mathbf{A}\mathbf{R}(t+m)]^m = \mathbf{P}\mathbf{D}^m\mathbf{P}^{-1} \quad (9)$$

And $\underline{\alpha}(t+m)$ can be computed directly from $\underline{\alpha}(t)$ without computing intermediate $\underline{\alpha}$s at t + 1, t + 2,..., and t + m - 1. That is,

$$\begin{aligned}\underline{\alpha}(t+m) &= \underline{\alpha}(t)[\mathbf{A}\mathbf{R}(t+m)]^m \\ &= \underline{\alpha}(t)\mathbf{P}\mathbf{D}^m\mathbf{P}^{-1}\end{aligned} \quad (10)$$

It seems that obtaining the matrices, P and P^{-1}, would require time-consuming computation, especially when the dimension of the

matrix is large. This, however, is not so in this case, since efficient methods exist when [AR(t + m)] is upper triangular.

Training Algorithms

In the previous section, an efficient way of computing αs without computing the intermediate values when a long run of the same codewords are observed, was shown. βs can also be computed in the same way. In this section, two different training methods are introduced in which the same method is employed to efficiently perform the re-estimation. The first one, denoted as "Algorithm 1," is strictly based on the Baum-Welch re-estimation algorithm, while the second one, denoted as "Algorithm 2," is a slightly different version that performs better.

Algorithm 1

In the Baum-Welch re-estimation algorithm, the estimates of a_{ij} and $b_j(v)$, denoted as $\hat{a}_{ij}$ and $\hat{b}_j(v)$ respectively, are updated at each iteration based on the previous estimates as follows:

$$\hat{a}_{ij} = \frac{\gamma_{ij}}{\gamma_i} \tag{11}$$

$$\hat{b}_j(v) = \frac{\sum_{t \in O(t)=k} \alpha_j(t)\beta_j(t)}{\sum_{t=1}^{F} \alpha_j(t)\beta_j(t)} \tag{12}$$

in which

$$\gamma_{ij} = 1/p \sum_{t=1}^{F-1} \alpha_j(t)\hat{a}_{ij}\hat{b}_j(O[t+1])\beta_j(t+1) \tag{13}$$

$$\gamma_i = \sum_{j=1}^{n} \gamma_{ij} \tag{14}$$

Consider the computation of γ_{ij}. If $O(k+1) = O(k+2) = \ldots = O(k+m)$, then $b_j(O[k+1]) = b_j(O[k+2]) = \ldots = b_j(O[k+m])$. Thus the partial summation of Equation (13) for $k \le t \le k+m-1$, denoted as $\gamma_{ij}(k, k+m-1)$, can be written as

$$\gamma_{ij} = 1/p\ [a_{ij}b_j(O[k+1])] \sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) \tag{15}$$

Computation of Equation (16) directly requires $\alpha_i(t)$ and $\beta_j(t+1)$ to be computed at $t = k, k+1, \ldots, k+m-1$. With a different strategy, which is shown below, great gains in efficiency can be achieved, especially when m is large. First express $\underline{\alpha}(t)$ and $\underline{\beta}(t+1)$ for $k \le t \le k+m-1$ in terms of $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$ as follows:

$$\underline{\alpha}^T(t) = \underline{\alpha}^T(k)[\mathbf{AR}(k+1)]^{t-k} \tag{16}$$

$$\underline{\beta}(t+1) = [\mathbf{AR}(k+1)]^{m-t+k-1}\,\underline{\beta}(k+m) \tag{17}$$

It follows that

$$\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) = \sum_{t=k}^{k+m-1} [\underline{\alpha}(t)\underline{\beta}^T(t+1)]_{ij} \tag{18}$$

$$\sum_{t=k}^{k+m-1} [(\{AR\}^{t-k})^T\,\underline{\alpha}(k)\underline{\beta}^T(k+m)(\{\mathbf{AR}\}^{m-t+k-1})^T]_{ij}$$

in which $(*)_{ij}$ denotes i-j component of matrix (*) and R = R(k+1) for simplicity. As shown in the previous section, AR can be decomposed so that AR = PDP^{-1}. Then Equation (19) can be rewritten as follows:

$$\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1)$$

$$= \sum_{t=k}^{k+m-1} [\mathbf{P}^{-T}\mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1}\mathbf{P}^T]_{ij} \tag{19}$$

$$= [\mathbf{P}^{-T}(\sum_{t=k}^{k+m-1} \mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1})\mathbf{P}^T]_{ij}$$

If

$$\hat{\underline{\alpha}}(k) = \mathbf{P}^T\underline{\alpha}(k) \tag{20}$$

$$\hat{\underline{\beta}}^T(k+m) = \underline{\beta}^T(k+m)\mathbf{P}^{-T} \tag{21}$$

then Equation (20) can be written more neatly so that

$$\sum_{t=k}^{k+m-1} \alpha_i(t)\beta_j(t+1) = [\mathbf{P}^{-T}\mathbf{M}\mathbf{P}^T]_{ij} \tag{22}$$

in which

$$\mathbf{M} = \sum_{t=k}^{k+m-1} \mathbf{D}^{t-k}\mathbf{P}^T\underline{\alpha}(k)\underline{\beta}^T(k+m)\mathbf{P}^{-T}\mathbf{D}^{m-t+k-1} \tag{23}$$

$$= \sum_{t=k}^{k+m-1} \mathbf{D}^{t-k}\hat{\underline{\alpha}}(k)\,\hat{\underline{\beta}}^T(k+m)\mathbf{D}^{m-t+k-1}$$

Now consider the computation of M. The $(i-j)^{th}$ component of M, M_{ij}, can be expressed as

$$M_{ij} = \sum_{t=k}^{k+m-1} d_i^{t-k}\,\hat{\alpha}_i(k)\,\hat{\beta}_j(k+m)\,d_j^{m-t+k-1} \tag{24}$$

$$= (\hat{\alpha}_i(k)\,\hat{\beta}_j(k+m)) \sum_{t=k}^{m-t+k-1} d_i^{t-k}d_j^{m-t+k-1}$$

Since it was assumed that $d_i \neq d_j$ if $i \neq j$, the summation can be reduced so that

$$\sum_{t=k}^{k+m-1} d_i^{t-k} d_j^{m-t+k-1} = \begin{cases} \dfrac{d_j^m - d_i^m}{d_j - d_i} & \text{for } i \neq j \\ m(d_i)^{m-1} & \text{for } i = j \end{cases} \tag{25}$$

Thus

$$M_{ij} = \begin{cases} \dfrac{d_j^m - d_i^m}{d_j - d_i} \; \hat{\alpha}_i(k)\hat{\beta}_j(k+m) & \text{for } i \neq j \\ m(d_i)^{m-1} \; \hat{\alpha}_i(k)\hat{\beta}_j(k+m) & \text{for } i = j \end{cases} \tag{26}$$

In summary,

$$\gamma_{ij}(k, k+m-1) = 1/p[\mathbf{P}^{-T}\mathbf{M}\mathbf{P}^{T}]_{ij}(a_{ij}b_j\{O(k+1)\}) \tag{27}$$

It is worth noting that only the upper triangular portion of M must be computed, since only $\gamma_{ij}(k, k+m-1)$ is needed for $i \leq j$, and the matrices, P^{-T} and P^{T}, are lower triangular.

Second, consider the numerator of Equation (12) for the re-estimation of $b_j(v)$. Under the same assumption that $O(k+1) = O(k+2) = \ldots = O(k+m) = v$, the partial summation of the numerator, $\Sigma_{t \varepsilon O(t)=v} \alpha_i(t)\beta_j(t)$, for $k+1 \leq t \leq k+m$ can be expressed in terms of $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$,

$$\sum_{t=k+1}^{k+m} \alpha_j(t)\beta_j(t) = \sum_{t=k+1}^{k+m} (\underline{\alpha}[t]\underline{\beta}^{T}[t])_{jj} \tag{28}$$

$$= [\mathbf{P}^{-T} \sum_{t=k+1}^{k+m} (\mathbf{D}^{t-k}\mathbf{P}^{T} \underline{\alpha}(k)\underline{\beta}^{T}(k+m)\mathbf{P}^{-T}\mathbf{D}^{k+m-t})\mathbf{P}^{T}]_{jj}$$

Equation (26) is not unlike Equation (20) and can be evaluated similarly. In fact, if

$$\hat{\mathbf{M}} = \sum_{k+1}^{k+m} \mathbf{D}^{t-k}\mathbf{P}^{T}\underline{\alpha}(k)\underline{\beta}^{T}(k+m)\mathbf{P}^{-T}\mathbf{D}^{k+m-t} \tag{29}$$

It can be observed that $\hat{\mathbf{M}}$ is the product of D and M, that is,

$$\hat{\mathbf{M}} = \mathbf{D}\mathbf{M} \tag{30}$$

Hence, once M is obtained to compute $\gamma_{ij}(k, k+m-1)$, Equation (26) can be computed with only a few more operations as follows:

$$\sum_{t=k+1}^{k+m} \alpha_j(t)\beta_j(t) = [\mathbf{P}^{-T}\mathbf{D}\mathbf{M}\mathbf{P}^{T}]_{jj} \tag{31}$$

It should also be noted that in the partial summations involved for the re-estimations of a_{ij} and $b_j(v)$, $\underline{\alpha}$s and $\underline{\beta}$s are not required to be computed at every time unit. For example, if we consider the assumption given above that $O(t+1) = O(t+2) = \ldots = O(t+m)$, only $\underline{\alpha}(k)$ and $\underline{\beta}(k+m)$ are required in the partial summations, that is, all the intermediate $\underline{\alpha}$s and $\underline{\beta}$s do not have to be computed.

Algorithm 2

The algorithm presented here can be considered as a sampled version of Baum-Welch re-estimation algorithm. Unlike the Baum-Welch algorithm, which is formulated by Equations (11) and (12), in the new algorithm, only samples of γij are used. Equation (13) can be rewritten as follows:

$$\gamma_{ij} = \sum_{t=1}^{F-1} \gamma_{ij}(t) \tag{32}$$

in which

$$\gamma_{ij}(t) = 1/p\ \alpha_i(t)\, \bar{a}_{ij} \bar{b}_j(O[t+1])\beta_j(t+1) \tag{33}$$

The re-estimation equations (11) and (12) can also be written in terms of $\gamma ij(t)$

$$\bar{a}_{ij} = \frac{\sum_{t=1}^{F-1} \gamma_{ij}(t)}{\sum_{j=1}^{n} \left(\sum_{t=1}^{F-1} \gamma_{ij}(t)\right)} \tag{34}$$

$$\bar{b}_j(v) = \frac{\sum_{t \& o(t)=v} \left(\sum_{i=1}^{n} \gamma_{ij}(t-1)\right)}{\sum_{t=1}^{F} \left(\sum_{i=1}^{n} \gamma_{ij}(t-1)\right)} \tag{35}$$

In the new algorithm, $\gamma ij(t)$ is sampled at every k^{th} time unit. This is similar to assuming that

$$\begin{aligned} &\gamma ij(1) = \gamma ij(2) = \ldots = \gamma ij(k) \\ &\gamma ij(k+1) = \gamma ij(k+2) = \ldots = \gamma ij(2k) \\ &\gamma ij(2k+1) = \gamma ij(2k+2) = \ldots = \gamma ij(3k) \\ &\quad \vdots \end{aligned} \tag{36}$$

Using this and the assumption that $F = mk$ for some integer m, Equation (31) becomes

$$\bar{a}_{ij} = \frac{\sum_{r=0}^{m-1} \gamma_{ij}(\tau k+1)}{\sum_{j=1}^{n} \sum_{r=0}^{m-1} \gamma_{ij}(\tau k+1)} \tag{37}$$

Although this algorithm is not guaranteed mathematically to converge, it has never been observed in practice to be a problem. Not only is it computationally simpler than Algorithm 1, but it also resulted in improved performance.

Front End Analysis

The approach adopted is based on a linear model of speech that is time-invariant over short intervals. This is the traditional model often used in speech recognition and coding applications. However, natural smooth changes occurring in the system, as well as additive uncorrelated noise, that are allowed for the linear model may also have explicit modeling of time-varying system parameters. Since many phonemes are characterized by a

particular evolution in time rather than by steady state or target spectra, this model is more powerful than more traditional ones. In particular

$$\begin{cases} \mathbf{X}(k) = \Phi(k)\mathbf{X}(k-1) + \Gamma(k)w(k) \\ s(k) = \mathbf{H}^T\mathbf{X}(k) + v(k) \end{cases} \tag{38}$$

in which the vector $X(k) = [x(k)x(k-1)\ldots x(k-p+1)]^T$, $x(k)$ is the speech without noise, $w(k)$ the noise input and $\Gamma(k)$ its gain, $H^T = [1,0,0,\ldots,0]$, $v(k)$ the additive noise, and $\Phi(k)$ characterizes the time-varying vocal tract filter.

Systems similar to this have been used to model many varied signals arising in sonar, heart monitoring, aircraft control, and so forth. In the linear prediction synthesis, model $\Phi(k)$ remains constant over 10- to 30-millisecond intervals, and $v(k)$ is zero. In the LPC analysis model, $v(k)$ is generally assumed to be zero so that $\Phi(k)$ can be estimated every 10 to 30 milliseconds. Recursive linear least squares estimation based on the model falls within the general area of Kalman filtering, which allows one to efficiently compute the least squares estimate of $X(k)$ from the least squares estimate of $X(k-1)$ and $s(k)$. If the system has been modeled correctly, the property to exploit is the prediction error, $\varepsilon(k)$, which would be white, and it should have a predictable ratio of its power to the unfiltered signal's power. If there are L possible models from which the observed signals were generated, this idea can be used for computing the relative likelihood of each model, given the observed signal. In the following, the front end process is explained in detail on the Kalman filtering process followed by the decision-making process.

Kalman Filtering

In the Kalman filtering process, there are L distinct competing models, each of which has the form

$$\begin{cases} \mathbf{X}(k) = \Phi\mathbf{X}(k-1) + \Gamma(k)w(k) \\ s(k) = \mathbf{H}\mathbf{X}(k) + v(k) \end{cases} \tag{39}$$

in which $\Phi = \begin{bmatrix} a(1) & a(2) & \ldots & a(p) \\ 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}$

$$\mathbf{H} = [1,0,0,\ldots,0]$$

$$E[v(k)] = 0, \quad E[v(k)v(l)] = \sigma_v^2(k)\delta_{kl}$$

$$E[\mathbf{w}(k)] = [0,0,\ldots,0]^T$$

$$E[\mathbf{w}(k)]\mathbf{w}(l) = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix} \delta_{kl}$$

$$\Gamma(k) = \begin{bmatrix} g(k) & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 0 \end{bmatrix}$$

and $a(1),a(2),\ldots,a(p)$ are linear prediction coefficients that characterize the model. This model results in the following time-recursive formula which gives the linear least squares estimate of $X(k)$ given $s(k-1)$, $s(k-2),\ldots,s(0)$.

$$\varepsilon(k) = s(k) - \mathbf{H}\hat{\mathbf{X}}(k|k-1) \tag{40}$$

$$\sigma_\varepsilon^2(k) = \mathbf{H}\mathbf{P}(k|k-1)\mathbf{H}^T + \sigma_v^2(k) \tag{41}$$

$$\mathbf{M}(k) = 1/\sigma_v^2 \; \mathbf{P}(k|k-1)\mathbf{H}^T \tag{42}$$

$$\hat{\mathbf{X}}(k|k) = \hat{\mathbf{X}}(k|k-1) + \mathbf{M}(k)\varepsilon(k) \tag{43}$$

$$\hat{\mathbf{X}}(k+1|k) = \Phi\hat{\mathbf{X}}(k|k) \tag{44}$$

$$\mathbf{P}(k|k) = \mathbf{P}(k|k-1)\mathbf{M}(k)\mathbf{M}^T(k)\,\sigma_v^2 \tag{45}$$

$$\mathbf{P}(k+1|k) = \Phi\mathbf{P}(k|k)\Phi^T + \Gamma(k)\Gamma^T(k) \tag{46}$$

in which $\varepsilon(k)$ is the innovations sequence, $\sigma_\varepsilon^2(k)$ the variance of the innovations, $\mathbf{M}(k)$ the Kalman gain, and $\mathbf{P}(k|r)$ the covariance of the estimate error $\mathbf{X}(k) - \hat{\mathbf{X}}(k|r)$. The initial condition is given as follows:

$$\begin{cases} \hat{\mathbf{X}}^T(0|0) = [s(0)s(-1)\ldots s(-p+1)] \\ \mathbf{P}(0|0) = \sigma_v^2(0)\mathbf{I} \end{cases} \tag{47}$$

With the innovations sequence obtained from each model, a likelihood test is performed in a recursive manner. If $\varepsilon_i(k)$ is denoted the innovation produced by model i at time k and pi(k) the probability that model i generates s(k), then

$$p_i(k) = \frac{N(\varepsilon_i[k], \sigma_{\varepsilon_i}^2[k])p_i(k-1)}{\sum_{j=1}^{L} N(\varepsilon_j[k], \sigma_{\varepsilon_j}^2[k])p_j(k-1)} \tag{48}$$

in which $\sigma^2_{\varepsilon_j}(k)$ is the variance of $\varepsilon_j(k)$ when model j is correct, and N(a,b) represents the gaussian density of zero mean with the variance b evaluated at a. The model with the largest p is then chosen.

Experiments

A set of recognition experiments was performed with clean speech, noisy speech of SNR = 26 dB, and of SNR = 20 dB. The isolated words used in the experiments are "break," "change," degree," "eight," "eighty," "enter," "fifty," "fix," six," "go." Each word has 12 utterances, 6 of which were used for training the HMMs. Each utterance was passed through the Kalman-filtering process with three different levels of white gaussian noises as stated above, which produced three different sets of code words, one for clean speech, one for the noisy speech of SNR = 26dB, and one for the noisy speech of SNR = 20dB. In the Kalman-filtering process, the variances of the generating noise and the additive noise were updated every 80 samples, and the initial conditions were reset accordingly at the same time. The filter order was 14 for each of the 64 different filters.

1. With clean speech: two errors of 120 = 1.7%

   One error from the set used for training: "six" recognized as "fix"
   One error from the set not used for training: "eight" recognized as "eighty"

2. With noisy speech of SNR = 26dB: eight errors of 120 = 6.7%

   Zero errors from the set used for training,
   Eight errors from the set not used for training: "eight" recognized as "eighty" (4), "fix" recognized as "six" (1), "six" recognized as "fix" (3)

3. With noisy speech of SNR = 26dB and clean speech: the recognition of noisy speech. Six errors of 120 = 5%

   Zero errors from the set used for training
   Six errors from the set not used for training: "eight" recognized as "eighty" (3), "fix" recognized as "six" (1), "six" recognized as "fix" (2);

   the recognition of clean speech: seven errors of 120 = 5.8%
   Two errors from the set used for training: "eight" recognized as "eighty," and "six" recognized as "fix"
   Five errors from the set not used for training: "eight" recognized as "eighty" (5)

It is interesting to note that the models trained with both clean and noisy speech give higher recognition rates for noisy speech (compare the results of 2 and 3) than the ones trained with only noisy speech, while giving a lower recognition rate for clean speech (compare the results of 1 and 3) than the ones trained with only clean speech. It may be interpreted as clean speech giving positive information for the training of noisy speech models, and noisy speech giving negative information for the training clean speech models. This behavior has been observed on several occasions. This method has a base line performance better than that of the previous section. Although it will not operate at really low SNRs, if it is in the range of 20 dB, excellent performance results.

Recognition of Speech in Noise Using Front End and Distance Measures

In this portion of the study, no enhancement of the speech was performed before recognition. Instead, a front end and a distance measure, which were more robust in noise than conventional methods, were examined. The front end was based on (mel)cepstral and delta-(mel)cepstral parameters. The distances were based on a projection measure designed to minimize noise.

The class of projection measures used in this study is based on theoretical and empirical observations found by Mansour and Juang (1988). From a theoretical investigation of several speech models, including the additive power spectrum, constrained autoregressive moving average (ARMA), and pole displacement models for speech plus noise, it was shown that the norms of truncated cepstral sequences were reduced in the presence of white gaussian noise. In addition, from studying histograms, Mansour and Juang (1988) made three important observations:

1. For a given SNR, cepstral vectors with larger norms were less affected than vectors with lower norms.

2. Lower order cepstral coefficients were affected more than higher ones.

3. The direction of the cepstral vector was less affected by noise contamination than the vector norm (never exceeding 90°).

From these observations, a family of distortion measures was formulated and used in a dynamic time warping (DTW) recognition scheme for recognition of noisy speech. In the present study, this set of distortion measures is further developed for use in a continuous density HMM recognition system. In addition, the measure is used with an augmented feature vector consisting of not only the cepstral coefficients but also a set of time differential (delta) coefficients.

Projection Measure Computation

To compensate for the norm degradation observed, an equalization factor is needed to account for the shrinkage in the norms. Such a scale factor would effectively remove the cepstral norms from the distance computation. This results in a Euclidean distance measure between test and reference speech spectra, $f_t(e^{j\omega})$ and $f_r(e^{j\omega})$, respectively, of the following form:

$$d_{\text{Euclidean}}(f_t, f_r) = (C_t - \lambda C_r)^T (C_t - \lambda C_r) \qquad (49)$$

in which C_t and C_r are the N^{th} order cepstral vectors of the test and reference speech signals, respectively, and λ is the equalization factor which is an implicit function of the SNR. The optimal value of λ which minimizes the above distance can easily be found by applying the orthogonality principle and is equivalent to the projection of C_t onto C_r is

$$\lambda_{opt} = \frac{C_t^T C_r}{C_r^T C_r} \qquad (50)$$

Hence, the compensated distance measure becomes the Euclidean distance between the test (noise-degraded speech) cepstral vector, C_t, and its projection onto the reference (noise-free speech) cepstral vector, C_r:

$$d_{proj}(f_t, f_r) = (Ct - P_r C_t)^T (C_t - P_r C_t) \quad (51)$$

in which P_r is the projection operation matrix to the vector space C_r:

$$P_r = \frac{C_r C_r^T}{C_r^T C_r} \quad (52)$$

In this project, the projection measure was extended to be used in an HMM speech recognition system. In the context of a continuous observation HMM recognizer, the Euclidean distance is weighted by the inverse covariance matrix of each state density, called the inverse covariance-weighted Euclidean distance:

$$d_{stand}(f_t, f_r) = x^T C^{-1} x - b^T C^{-1} b \quad (53)$$

in which $f_t(e^{j\omega})$ is the test spectrum, $f_r(e^{j\omega})$ is the probability density of the HMM state, x is the cepstral coefficient vector of the test observation, b is the HMM state density mean vector, and C is the state covariance matrix. Modifying this weighted Euclidean measure, the weighted projection measure for use in an HMM recognition system becomes

$$d_{proj}(f_t, f_r) = x^T C^{-1} x - \frac{x^T C^{-1} b b^T C^{-1} x}{b^T C^{-1} b} \quad (54)$$

in which the variables are as described above and the second term is the projection of the test x onto b weighted by the inverse covariance matrix C^{-1}. This was the measure used in this study and was compared with the standard weighted Euclidean distance described in Equation (53) in recognition experiments.

Recognition System

Feature Extraction

This study used the MIT Lincoln Labs speech data base with one general American male speaker and a confusable vocabulary of 34 words ("eight","eighty"; "wide","wide"; "go","no","oh"; etc.). Five noise-free tokens are used for training the HMMs, while five noise-degraded tokens of each word are used for testing. The data base was recorded at a sampling rate of 16 kHz and then filtered and down-sampled to 8 kHz. An 8th order LPC analysis was performed on frames of speech 45 msec long with a Hamming window every 15 msec, producing an overlap of 66%. The words were endpointed by applying an appropriate energy threshold. From the LPC filter coefficients, two types of coefficients were computed, the cepstral and melcepstral coefficients. In addition, a set of time differential coefficients were used to describe the time-varying characteristics of the speech wave form. Various numbers of coefficients were tried, with 12 to 16 coefficients performing equally well in recognition tasks. Hence, a 32nd order feature vector consisting of 16 coefficients and 16 delta-coefficients was used for each frame of speech data.

1. Cepstral Coefficients

Sixteen coefficients (excluding the c(0) term) were recursively computed from the 8th order LPC filter coefficients:

$$-jc(j)-ja(j) = \sum_{i=1}^{j-1} (j - i)c(j - i)a(i), \quad j = 1,2,\ldots,16, \qquad (55)$$

in which the c(j)s are the cepstral coefficients, the a(j)s are the LPC filter coefficients, and a(j) = 0 for i > p is assumed. A set of 16 time differential (delta) coefficients is computed from a 7-point Parks-McClellan finite impulse response (FIR) differentiator filter (with less than 5% maximum deviation).

2. Melcepstral Coefficients

Sixteen coefficients (excluding the 0^{th} term) were computed from the 25-point LPC power spectrum filtered by a bank of 24 triangular windows spaced linearly below 1 kHz and logarithmically from 1 kHz to 4 kHz:

$$mfcc(i) = \sum_{k=1}^{nfilt} \log X_k \cos \left[i(k - 1/2)\ \pi/nfilt \right] \qquad (56)$$

in which nfilt is the number of mel-spaced filters, i = (1,2,...,16), and X_k is the energy output of the k^{th} filter. As with the cepstral coefficients, a set of 16 delta-melcepstral coefficients is also computed.

The motivation for using the melcepstral coefficients is that they are perceptually based, modeling the frequency resolution of the human ear derived from tuning curves, which tend to be linear below 1 kHz and logarithmic above 1 kHz. This suggests that they would be superior for speech recognition in noise, since humans have little trouble discerning speech from background noise. In addition, they are close to the principal component decomposition of the speech spectrum (Pols, van der Kamp, & Plomp, 1969) and have been found to be well correlated with perceptual judgment data (Pols, 1971).

Several numbers of mel-spaced windows were tried, ranging from 12 to 24 windows. Above roughly 20 windows, performance was not significantly improved with each added window. Hence, the value of 24 windows was settled upon as a reasonable number needed to capture the perceptually significant energy bands. The same number of mel-spaced or critically spaced windows has been suggested by O'Shaughnessy (1987) to model the human hearing to a first order.

Unlike the cepstral coefficients, the melcepstral coefficients did not show norm degradation at very low SNRs (typically affecting the fricatives). This is because of the over-emphasis on the higher frequencies by the log-spaced windows in the melcepstral coefficient computation. To compensate for this over-weighting, various weightings were tried. One such method is normalizing each mel-spaced window energy output by the area of the window or effectively just the window length. This results in a set of normed melcepstral coefficients which are also investigated:

$$mfcc\_norm(i) = \sum_{k=1}^{nfilt} \log X_k/L_k \cos \left[i(k - 1/2)\ \pi/nfilt \right] \qquad (57)$$

in which X_k is the energy output of the kth mel-spaced filter and L_k is its length. Like the other coefficients used, a set of delta coefficients was used with the normed melcepstral coefficients. Note that such a normalization does not affect the Euclidean distance between two speech signals since the constants involving the L_ks will cancel. However, this is not the case with the projection measure and thus, such a normalization is used.

HMM Recognition System

A discrete state, continuous density HMM recognition system was used. The forward-backward training algorithm was used to generate the 10 state left-to-right (with no skip transitions) models for each word described by the probabilistic parameter set (π, A, B). The initial state probability vector π determines the starting state in the model and was set to π = (1,0,...,0), allowing only starting in state 1. The transition matrix A was initialized with diagonal terms (self transitions) set to 0. and the off-diagonal (transitions to the next adjacent state set to 0.2, with all other transitions prohibited. The matrix is randomly perturbed so each model was begun at a different point.

The B parameter set represents the probability density function in each state of the model. The B densities were initialized for the algorithm by segmenting each training token into 10 equal parts and having this be the initial estimate of the state sequence. From this estimate, a (mel)cepstral mean vector and covariance matrix were extracted to describe the gaussian observation probability function in each state. In addition, the mean and covariance matrix for a set of delta-(mel)cepstral vectors was also determined. Because of problems of inefficient training data (only five tokens per word model), the pooled or grand variance matrix of all the training data was used for each state density and a large number of states (10) was needed.

With the probabilistic set (π, A, B) initialized as described above, the forward-backward algorithm was then used to rain the HMMs. At each iteration, the algorithm determined the probability of all possible state paths through the current model estimate using the set of training tokens. From these probabilities, the transition matrix A was re-estimated. However, the initial state vector π was not re-estimated. Also, these probabilities were used to weight he present observation parameter vectors as they were averaged together to form the new estimate of the state mean vectors. Since the covariance matrix was assumed to be universal, it was not re-estimated in the procedure. Typically, three to four iterations were needed to achieve model convergence.

It is important to note that all HMM training was done on the noise-free speech and only the covariance-weighted Euclidean distance was used, not the projection measure. Two sets of training were done: one with both the (mel)cepstral and delta coefficients used together as the feature vector for each frame of speech and the other with only the (mel)cepstral coefficients used as the feature vector.

For recognition, the 'Viterbi algorithm was used to determine the most probable path through the word models. A white gaussian noise generator was used to add noise to five tokens of each vocabulary word for various SNRs. These noise-degraded words were then tested using the recognition system trained on the noise-free speech. Both the new projection measure and the standard inverse covariance-weighted Euclidean distance were used in the comparison stage.

Recognition Results

Recognition experiments were conducted to evaluate the effectiveness of three factors: the projection measure, the inclusion of delta-parameters, and the melcepstral representation. Tables 7 and 8 report experimental recognition results for various levels of added white gaussian noise obtained using the cepstral coefficients, melcepstral coefficients, and a set of melcepstral coefficients computed by normalizing the energy output of each triangular window by its window length. Experiments were conducted for feature observations with and without augmenting the feature vector with delta-coefficients. Table 7 lists results using the inverse covariance-weighted Euclidean distance, while Table 8 lists results using the inverse covariance weighted projection measure.

Table 7

Recognition Results Using the Inverse Covariance-weighted Euclidean Distance
(Results are given for feature vector without delta coefficients [nd]
and with delta coefficients [d]. Units are in percentages.)

Type of coefficients	Noise free	20 dB	15 dB	10 dB	5 dB
cep/nd	97.1	70	41.8	24.7	13.5
mel/nd	95.3	54.1	25.9	11.2	4.1
norm-mel/nd	95.3	54.1	25.9	11.2	4.1
cep/d	97.1	82.4	59.4	35.9	17.1
mel/d	97.7	68.8	49.4	21.8	5.9
norm-mel/d	97 7	68.8	49.4	21.8	5.9

As can be seen from Tables 7 and 8, the weighted projection measure significantly improved recognition performance for all parameter representations and over all SNRs. Even at a low SNR of 5 dB, improvements from only 5.9% recognition accuracy (basically chance) to a respectable 51.2% accuracy were achieved. In addition, the inclusion of delta-parameters also noticeably improved recognition accuracy, with increases of 10 to 23 percentage points. Such results indicate the value of time-differential features for speech recognition in noise.

Table 8

Recognition Results Using the Weighted Projection Measure
(Results are given for feature vector without delta coefficients (nd) and with delta coefficients (d). Units are in percentages.)

Type of coefficients	Noise free	20 dB	15 dB	10 dB	5 dB
cep/nd	96.5	86.5	76.5	57.6	32.4
mel/nd	95.3	85.6	76.5	59.4	35.3
norm-mel/nd	95.6	87.1	76.5	57.6	40.0
cep/d	95.3	95.3	90.0	70.6	45.3
mel/d	97.1	96.5	88.8	75.3	44.1
norm-mel/d	97.1	97.1	93.5	80.0	51.2

While not performing well using the weighted Euclidean distance, the melcepstral representation showed the greatest improvement in performance using the projection measure. The normed melcepstral representation consistently out-performed the cepstral representation. It also out-performed the un-normalized melcepstral representation suggesting that the window energy normalization helps lessen the effects of broadband noise in the higher frequency bands. Such results suggest further that the projection measure somehow de-emphasizes this greater mismatch that is found in the higher frequency bands in the melcepstral representation.

Conclusions

As shown, three factors were found to enhance recognition accuracy in the presence of white noise:

1. Using the projection measure instead of the standard Euclidean distance.

2. Augmenting the feature vector with a set of time differential parameters.

3. Using a melcepstral instead of a cepstral representation for the speech (with the projection measure).

The projection measure is desirable for it can easily be incorporated into an HMM recognition system within the probability calculations. The measure is inherently independent of the type of noise used, so an estimate of the noise spectrum is not needed. This is an important consideration when training the system in noise-free environments and using the system in environments with varying degrees and types of noise where measures robust to noise are needed.

SPEECH IN STRESS

This portion of the report describes the data base compiled for the research, how the speech is characterized under various conditions, and how recognition can be improved using these data.

Data Base

A comprehensive stress data base, which has already been established at Georgia Tech, is partitioned into five domains encompassing a wide variety of stresses and emotions. A total of 32 speakers (13 female, 19 male), with ages ranging from 22 to 76 were employed to generate more than 16,000 utterances. Table 9 illustrates the various domains of the data base. Each domain is discussed separately. The reasons for including a particular domain along with its contributions are also indicated.

Table 9

The Georgia Tech Speech Under Stress Data Base

Speech Under Stress Data Base
Georgia Institute of Technology, School of Electrical Engineering

Domain	Type of stress or emotion	Number of speakers	Number of utterances	Source
Psychiatric analysis	Depression Fear Anxiety Anger	six female two male age ranges 34 to 76	600 (at present)	Emory University School of Medicine Dept. of Psychiatry
Talking style	Slow Fast Soft Loud Angry Clear Question	all male 3 General 3 New York 3 Boston	8,820 (total)	MIT Lincoln Labs Boston, MA 35 aircraft communication words
Single tracking task	Work load (moderate) (high) Lombard	all male	1,890 (total)	MIT Lincoln Labs calibrated work load tracking task
Dual tracking task	Work load (moderate) (high)	four female four male	4,320 (total)	Georgia Tech acquisition tracking compensatory tracking
Subject motion fear tasks	G-force Lombard Noise Fear Anxiety	three female four male	400 (total)	Georgia Tech controlled motion noisy environment

The analysis and identification of emotion is speech is an important facet for psychiatric analysis. A psychiatrist must be able to determine quickly and reliably the emotional state of a patient. The task of educating new psychiatrists to identify the emotional state of a patient is quite difficult. Therefore, researchers in this field have expressed the desire to possibly develop a systematic set of recordings which student psychiatrists could use to better understand how emotion is relayed. Another aim for this portion is the development of an objective measure of stress or emotion based on a set of composite parameters. Patients from Emory Medical University's Department of Psychiatry who underwent psychiatric analysis were recorded using a high quality microphone and tape recorder in a natural doctor-patient environment. Recordings from eight patients (six female, two male) were obtained. Each recording consists of approximately 1 hour of doctor-patient discussion. After the first few moments during an analysis session, all patients tended to disregard the microphone and continued without any inhibitions because of the recording equipment. Although the overriding emotion found during the recordings was mild to severe depression, brief passages of fear, anxiety, and/or anger have also been identified. The analysis of each recording involved subjectively marking phrases or individual words as being under stress. Each utterance was then low-pass filtered at 3.7 kHz and sampled at 8 kHz. For the first six recordings excellent examples were found of mild to severely depressed speech (in some cases, patients were even crying). Although the determination of each utterance was a subjective decision, the choice was usually quite clear. One slight problem is that direct comparison of phrases under stress and neutral conditions would not always be possible. This was solved by requesting that during the following analysis session, the patient would be instructed to repeat these words several times in a natural, relaxed frame, allowing for a more reliable analysis in the next section. A vocabulary of more than 600 words or phrases was collected.

The second portion of the Georgia Tech stress data base involved speech under various speaking styles. In an earlier investigation, Lippmann, Mack and Paul (1986) considered a multi-style training procedure for a traditional HMM, speaker-dependent, isolated work recognition system. By employing utterances spoken using various speaking styles, a larger sample space results for each utterance. Thus, when a word is presented to the recognition system, it has a higher probability of falling into the correct sample space since multi-style training will account for larger speech variations. For the system of Lippmann et al. (1986), overall recognition errors decreased form 20.7% to 9.8% using multi-style training. The data from this earlier investigation were obtained from MIT Lincoln Labs through personal contact with Richard Lippmann and Clifford Weinstein. This portion contains utterances under eight speaking styles (normal, slow, fast, soft, loud, question, clear enunciation, angry). The vocabulary includes 35 aircraft communication words containing a number of subsets that are difficult for recognition systems. These subsets include {go, hello, on, no}, {six, fix}, {white, wide, point}, {degree, three, thirty, freeze}, and {eight, eighty, gain, change}. The 35 words comprise a subset of 105 words presently used by Texas Instruments to evaluate recognition systems. The words were produced by nine male talkers sampling three major dialects (General American, Boston, New York). Each word was produced 28 times by each subject for a total of 8,820 words. In the present research effort, only the General American speakers will be considered; variations because of dialects will be addressed in future work. These words are already used and are well accepted in recognition research. In addition, the utterances were all produced in a quiet surrounding with little background interference.

The third and fourth portions involved speech produced while the subject was performing some task as a means of inducing work load. The purpose of speech recognition is to free the user so that outside tasks may be reformed. As indicated in the previous section, many researchers have tried to construct tasks ranging from timed mathematical tests to mild electrical shock as a way of generating stress. The intentions here are to focus on the types of stress associated in a speech recognition environment. Thus, in order to formulate this portion, the types of stress and work load tasks that might simulate the stress must first be discussed. As an example, speech recognition in a noisy aircraft cockpit might help a pilot better perform his duties. The particular stress experienced during flight are three-fold. The first of these is physical stress. This consists of abnormal conditions of air pressure, oxygen content low temperature, g-force, vibration, and so forth, that affect the pilot from the moment he leaves the ground in an unprotected cockpit. Such physical stress has already been recognized and investigated from previous research as a means for screening pilots from flying stress (Hanson & Wong, 1984). A second type of stress associated with such an environment is called *cognitive stress*. Cognitive refers to those intellectual processes of the brain concerned with information. This includes input from reality-based information (such as the pilot's surroundings), as well as that retrieved from memory files or created internally by fantasy. Cognitive stress in a flying situation is usually related to cockpit work load. When this becomes excessive, it affects the neurological state of the pilot (Nakatsui & Suzuki, 1970) and consequently his operational efficiency (Hanson & Wong, 1984). Finally, *affective stress* is produced whenever the input to consciousness (whether it be from an immediate awareness of external event, or a recollection of personally significant events of the past) is seen as threatening to the individual's safety, self esteem, or satisfaction of desires. Affective stress, in its initial phase, is not always harmful. For example, anxiety aroused by a sudden emergency can actually be beneficial in alerting the brain to optimal function. Intense affective stress, however, can seriously impair operational efficiency and can lead to severe mental or physical disablement in the form of psychoneurotic or psychosomatic disorders. The simulation of either physical or intense affective stress in a laboratory environment is not possible. However, some forms of cognitive or mild affective stress are possible with the help of interactive work load computer tasks.

The basic task of maintaining controlled flight requires several levels of psychomotor control. For example, a typical low altitude air-to-ground mission calls for the pilot to fly close to the terrain while navigating to a predetermined point, ascend to a higher altitude, acquire and track the intended target, take appropriate action, and finally to descend again close to the terrain for egress. During the time at high altitude, the aircraft is highly vulnerable to detection. Therefore, it is desirable that simulations mimic the tasks which are of great importance in target acquisition and tracking (the highly precise and coordinated manual aspects of flight control, and the relatively simple manual tasks involved in aiding the automated acquisition and tracking functions). The fact that in a single-seat aircraft, the pilot may be required to perform these tasks simultaneously is of particular concern in the generation of a suitable speech data base for the purposes of recognition in such an environment.

As indicated, the two types of tracking tasks considered are acquisition tracking and compensatory tracking. *Acquisition* or step tracking is a task where a sudden discrepancy between the target stimulus and a response marker must be nulled by the operator. Here, total acquisition time is typically divided into three segments: (1) reaction time--the time from the onset of

the discrepancy to the initiation of a control movement; (2) primary movement time--the duration of the first control movement, which usually nulls most of the discrepancy between the target and the marker; and (3) correction time--the time from the end of primary movement until the target and marker are in stable alignment. The second type of tacking task considered is compensatory. *Compensatory* tracking is a task where the response marker remains in a fixed position on the display. The target stimulus moves about in accordance with some input function which creates a discrepancy between the positions of the marker and the target. Thus, the term compensatory literally refers to the nature of the display. With a compensatory display, no preview of the to-be-tracked function is possible, although if the function is predictable, anticipation of the correct response may be possible. One problem with such a task is that the operator cannot readily distinguish between discrepancies that result from the input function and those that result from incorrect control movements unless the input function is fully predictable.

The third portion of the stress data base uses a work load task for stimulating stress originally proposed by Jex (1979). Jex formulated a set of standardized sub-critical tasks for tracking work load calibration. The important aspect of this work was that graded levels of mental work load were possible. In this approach, a single tracking task is developed which is the response of a marginally stable, single-pole system. The operator views a display of the error between the command input and plant output and corrects these with opposite pressure on a control stick. The degree of instability may be adjusted for varying degrees of difficulty. Two levels of work load difficulty were used in this section. Subjective ratings, performance data, and heart rate data indicated that the high work load (λ = 70%) was significantly more difficult than the moderate level (λ = 50%). This portion uses the same nine male talkers as in portion two of the data base, with a total of 1,890 words comprising this section. Again, only General American speakers were used.

In the fourth portion, the method proposed by Jex was initially considered in developing a dual task work load for induced stress. However, this approach represents a single compensatory tracking task, and in an environment such as an aircraft cockpit, the operator's responsibility would be much more demanding. Therefore, a dual tracking task that addresses pilots' two key goals (flight control and target acquisition) was considered. Task difficulty may be controlled by time constraints for completion or by increasing resource competition or motivation. By employing such a dual technique, it is now feasible to employ tasks that closely reflect the basic goals of the pilot. In this portion, a dual tracking task is considered which is similar to one previously developed by Folds, Garth, and Engelman (1986) for the United States Air Force (USAF) School of Aerospace Medicine. The primary tracking task is a pursuit task in which the input signal is determined by the sum of two sine functions. A constant is added and subtracted from the function to form two parallel sinusoids which are displayed, giving the appearance of a winding road which is scrolled down a computer screen over time. The response marker (output signal) is a small circle. The vertical position of the circle is fixed at the center of the display; the horizontal position is determined by the movement of a control stick in its X-axis. As the roadway moves downward, the operator's task is to move the control stick to position the circle as close to the center of the road as possible. After 20 seconds, a target acquisition task appears on the left portion of the screen. Here, two narrowly spaced vertical lines are drawn along with a small triangle. The vertical position of the triangle is fixed at the center of the display. A gaussian distributed random value is added to the triangle's horizontal position which moves it to the left or

right. The operator uses the X-axis movement of a second control stick to move the triangle back to the center of the two lines. These noise values are added at fixed times to the triangle's position, so the operator is not certain if movement at any time is a result of his actions or random movements. This represents the noise associated with an automatic target acquisition system which must be corrected by the pilot. After 40 seconds of performing both tasks, the primary compensatory task is disabled, leaving the secondary target task active for the final 20 seconds. Several parameters are available for increased task difficulty, the simplest being the overall time allowed to perform the task. Other possibilities exist, but variation in overall time constraint was chosen in order to keep the basic task structure similar to the previously used and accepted work load procedure.

To facilitate an organized manner for collection of an operator's speech, randomized words from the 35-word list used in portion two of the data base were displayed on the screen during the above three stages. Each word appeared in the same position, and the operator was instructed to read them quickly while performing the dual tracking task. Operators wore a high quality head-mounted directional microphone for recording. When a new word appeared on the screen, a low frequency tone was emitted. In some cases, a variation in the length of time a word remained on the screen was also used in order to generate higher stress levels. For the moderate stress level, every three-stage test lasted for 80 seconds. Each operator performed the three-stage dual work load task nine times, each with a different type of gaussian variation for the target acquisition task. A weighted root mean square (RMS) error was found for each task individually and combined. This was displayed between tests so that operators could observe how well they performed the tasks. Operators were instructed to give equal emphasis to all three tasks (pursuit task, target acquisition, and word entry). For the high work load case, the operators were required to perform all three tasks in half the time (40 seconds). A comparison of RMS errors indicated that the high work load case was significantly more taxing than the moderate case.

Portions 3 and 4 of the data base also include a small portion about speech produced in noise. It is known that talkers vary their speech characteristics when speaking in a noisy environment. For example, overall speech level as a function of external noise level has been shown to rise at the rate of 0.3 dB/dB noise to 1.0 dB/dB noise, depending on noise level and the specific task assigned to the speaker (Paul, 1987). This phenomenon, called the Lombard effect, was first noted in 1911 (Lippmann, Martin, & Paul, 1987). Speakers also tend to vary those factors related to speech clarity when presented with external noise. In both portions, noise was presented biaurally at an overall level of 85 dB sound pressure level (SPL).

The last portion of the stress data base uses two types of subject motion tasks. To simulate the sudden change in altitude or direction that might be experienced in an aircraft cockpit, two types of motion tasks were considered. (Originally, an aircraft flight simulator was considered. However, securing access and obtaining students knowledgeable in flight simulators or pilots willing to devote the time necessary to perform this proved futile.) Therefore, tasks had to be chosen which required little or no training, yet generated the type of stress (fear or anxiety) which might be experienced in an emergency situation. Two rides from Six Flags Over Georgia were chosen as suitable, the Scream Machine and Free Fall. The Free Fall ride lasts for about 60 seconds, with the free fall portion comprising about 10 seconds. Four seated passengers are strapped in an upright position into a car which is raised vertically to approximately 130 feet (10 stories). The car moves forward where it pauses for several seconds and then is released.

It drops vertically downward for about 100 feet, before rolling onto a horizontal portion of the track for deceleration. During the free fall portion, talkers repeated several prechosen words from the 35 word list (used in portions two through four of the data base). Speech was recorded using a high quality head-mounted directional microphone and cassette recording unit strapped to the talker's body.

The second motion task considered was the Scream Machine. This is a typical wooden frame roller-coaster which seats about 30 to 36 passengers. Because of the large number of passengers, higher levels of background screaming can be heard for those recordings. The overall ride consists of large vertical movements with small amounts of lateral movement during calm periods between drops. Initial tests gave little indication of variation in tape speed because of the motion of the recording equipment.

The entire ride lasts about 90 seconds. Talkers were instructed to say the word top when their car reached the top of a hill. Speakers repeated words from a word list card held in their hands (35 words from sections 2 through 4. Each speaker performed the task twice. Because of the increased task time for this ride, larger amounts of stressed utterances could be obtained. The speaker's location during the ride was identified based on timing and background noise. Only speech uttered during plunges in the ride were extracted for analysis. Figure 5 gives an overview of the ride and illustrates how each recording was partitioned and subjectively marked for stress with respect to time and position during the task. The chosen subjects were all native speakers of American English from the Georgia Tech community with no apparent speech deficiencies. Each talker (three female, four male) performed both subject motion tasks twice. A total of 400 utterances were identified as being under stress. In each subject motion task, at least four factors contributed to the type of speech recorded: g-force, background noise, Lombard effect, fear and/or anxiety.

The entire data base was collected with a clear expectation that only a portion of it could be analyzed during the course of the project. It is available, however, to any laboratory that desires it, and it will serve as a valuable tool for years to come.

Characterization of Speech in Stress

In this section, the various characteristics of stressed speech are described. The discussion is broken into two major categories: characteristics of the observed speech and characteristics of the glottal waveform. Some of the first set of characterizations are related to prosodics, which in many cases are the first indicators of stress, emphasis, or higher order meaning. The major components or this aspect include pitch, speaking rates, vowel durations, and intensity, as well as the variability of all the above. Another characterization relates to the formants, bandwidths, and their variabilities. The second set of characteristics is more subtle and more difficult to analyze. As discussed, a novel method of glottal extraction had to be developed to reach a satisfactory solution.

--- Speech Database ---
STRESS UNDER SUBJECT MOTION TASK

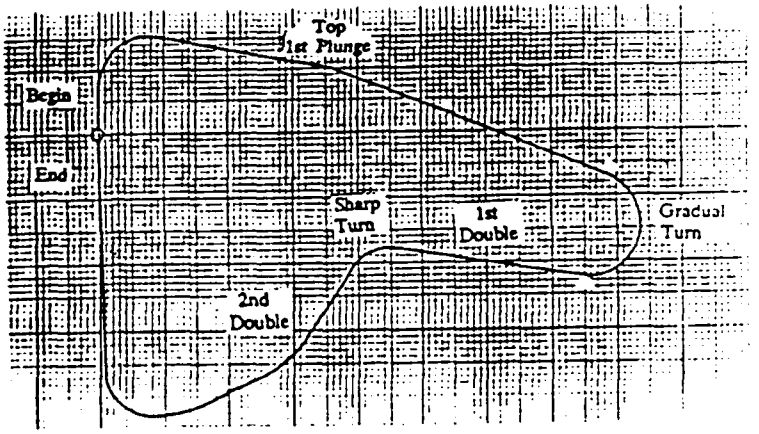
Performed: Nov. 2, 1986 John Hansen

Subject: *Joy Rose*

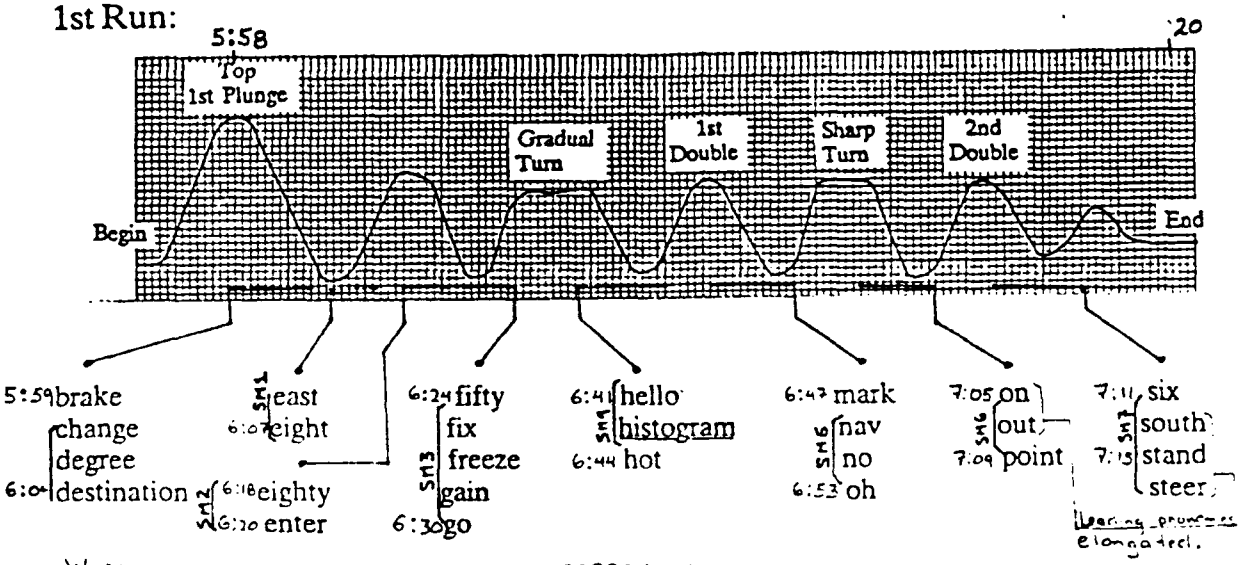
SM Profile & Location of Words:

-- Under Stress

-- VERY HIGH STRESS



1st Run:



2nd Run:

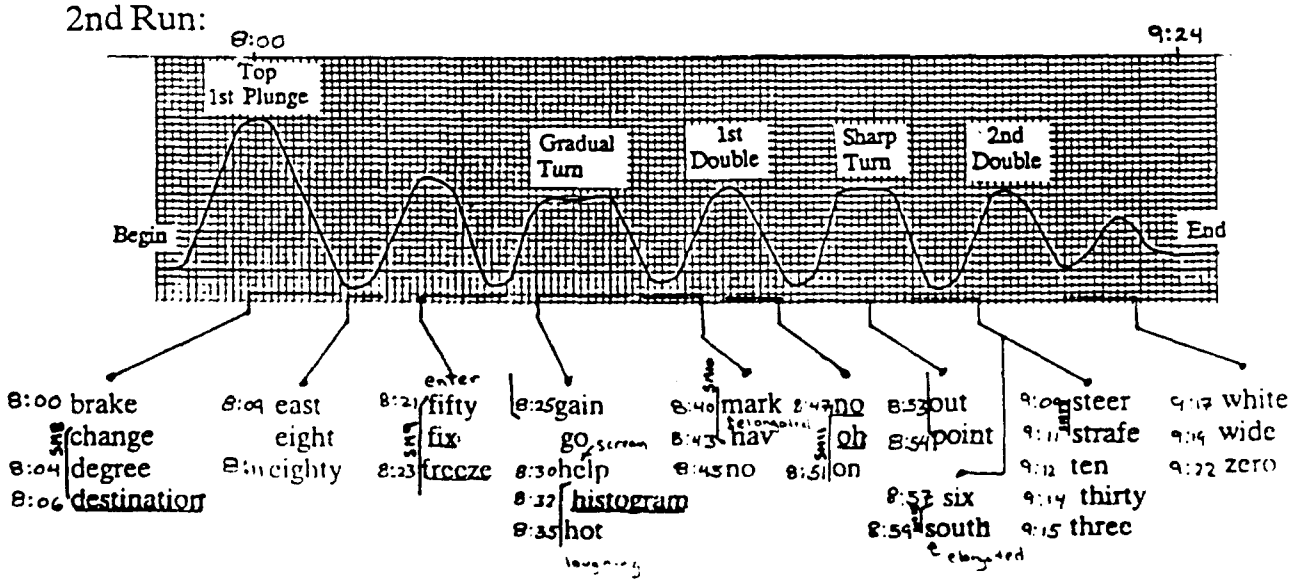


Figure 5. Profile of the scream machine motion task. (The position of utterances was marked based on timing, and those under stress were extracted for analysis.)

Non-glottal measures

Pitch

Measures of pitch and pitch variability were analyzed across the utterances in the data base. Table 10 describes the results. After statistical measures of significance (both parametric and non-parametric) were used, the following conclusions were made:

1. The position of mean pitch from highest to lowest versus speaking style is shown below. Results from both word lists and one of the many probe studies performed are included (sig. = significance with respect to neutral).

		Shift in mean pitch					
		Word List No. 2 385 utterances		Word List No. 1 385 utterances		Probe Study 55 utterances	
Pitch	Condition	(percent)		(percent)		(percent)	
Highest	Angry	+99	sig.	+73	sig.	+64	sig.
	Loud	+47	sig.	+48	sig.	+49	sig.
	Question	+44	sig.	+42	sig.	+42	sig.
	Lombard	+15	sig.	+10	sig.	+10	sig.
	Clear	+5	sig.	+4	sig.	+4	
	Fast	+5	sig.	+6	sig.	+5	
	Neutral						
	Slow	-2		+1		-2	
	Task Condition 70%	0		-2		-3	
	Task condition 50%	-2		-3		-2	
Lowest	Soft	-5	sig.	-5	sig.	-6	sig.

Analysis of fundamental frequency for Word List 1 over various peaking styles and stress conditions.

2. Mean pitch values may be used as significant indicators for speech in soft, fast, clear, Lombard, question, angry, or loud styles -. been compared to neutral conditions.

3. Loud, angry, question, and Lombard mean pitch are all significantly different from all other styles considered.

4. Mean pitch was not a significant indicator for moderate versus high task workload conditions.

5. Speech produced under the Lombard effect gave mean pitch values most closely associated with pitch from fast and clear conditions.

6. Variation in mean pitch appears to be a consistent and reliable stress indicator over a wide variety of conditions. The discriminating ability of mean pitch is clearly indicated in the Student's t-test significance tables.

Table 10

Analysis of Fundamental Frequency for Word List No. 1
Over Various Speaking Styles and Stress Conditions

Stress condition	Number points	Mean value	Max. value	Min. value	Average deviation	Standard deviation	Variation	Skew.	Kurtosis
Neutral	405	145.74	195.12	108.11	13.45	15.67	245.60	0.148	-0.770
Slow	759	147.57	195.12	105.26	16.19	18.67	348.63	0.087	-1.024
Fast	302	152.16	205.13	121.21	13.42	15.98	255.24	0.193	-0.578
Soft	390	138.81	266.67	119.40	6.76	11.23	126.04	4.964	46.501
Loud	593	216.31	307.69	97.56	42.15	49.02	2403.22	-0.509	-0.987
Anger	671	252.69	421.05	62.02	84.94	95.02	9028.02	0.122	-1.325
Clear	509	154.13	235.29	101.27	21.60	24.88	619.04	0.261	-0.910
Question	461	207.44	307.69	123.08	51.45	56.16	3153.75	0.231	-1.499
Cond50	379	141.46	200.00	112.68	13.56	16.13	260.33	0.414	-0.195
Cond70	384	142.49	186.05	105.26	13.64	15.96	254.61	0.151	-0.901
Lombard	506	160.49	242.42	100.00	21.10	24.36	593.20	-0.142	-0.796

Pitch Variability

1. The position of pitch variability from highest to lowest versus speaking style is shown below. Results from both word lists and one of the many probe studies performed are included (sig. = significance with respect to neutral).

Pitch	Condition	Shift in mean pitch Word List No. 2 385 utterances (percent)		Word List No. 1 385 utterances (percent)		Probe Study 55 utterances (percent)	
Highest	Angry	+506	sig.	+265	sig.	+527	sig.
	Question	+258	sig.	+264	sig.	+254	sig.
	Loud	+213	sig.	+186	sig.	+175	sig.
	Lombard	+55	sig.	+60	sig.	+50	sig.
	Clear	+59	sig.	+43	sig.	+49	sig.
	Slow	+19	sig.	-6		+3	
	Fast	+2		-9		+1	
	Task Condition 50%	+3		+5		-4	
	Task condition 70%	+2		+6		-5	
	Neutral						
Lowest	Soft	-28	sig.	-38	sig.	-58	sig.

2. Variance of pitch values may be used as significant stress indicators for speech in soft, loud, angry, clear, question, or Lombard styles when compared to neutral conditions.

3. Soft and loud pitch variance are significantly different from all styles considered.

4. Pitch variance was not significantly different for moderate versus high task work load conditions.

5. Pitch variance was unreliable for slow and fast stress conditions.

6. Pitch variance for clear and Lombard conditions are similar but different from all other styles considered.

Duration

Measures of duration and duration variability were analyzed across the utterances in the data base. Tables 11 and 12 describe the results.

Table 11

Average Word and Speech Class Duration Using Word List No. 1 (35 words/style, 11 styles = 385 words) Over Various Speaking Styles and Stress Conditions (in msec)

Stress condition	Average duration (msec)										
	Neutral	Slow	Fast	Soft	Loud	Anger	Clear	Question	Cond50	Cond70	Lombard
Word	478	827	353	509	650	662	666	526	482	501	572
Vowel	160	294	115	147	253	271	202	180	148	147	198
Consonant	71	107	52	87	73	62	128	74	79	86	73
Semi-vowel	60	126	57	71	76	85	83	84	71	68	97
Diphthong	192	374	147	210	294	315	199	209	176	178	249

Table 12

Variance of Average Word and Speech Class Duration Using Word List No. 1 (35 words/style, 11 styles = 385 words) Over Various Speaking Styles and Stress Conditions (in msec)

Stress condition	Variance of average duration (msec)										
	Neutral	Slow	Fast	Soft	Loud	Anger	Clear	Question	Cond50	Cond70	Lombard
Word	18.0	49.0	12.0	16.0	28.0	41.0	40.0	16.0	16.0	14.0	24.0
Vowel	7.9	21.0	3.6	6.2	19.0	23.0	17.0	9.0	7.6	7.6	13.0
Consonant	1.8	7.1	1.1	2.9	3.7	3.3	10.	2.9	2.5	3.3	2.6
Semi-vowel	0.7	7.1	1.0	1.3	2.9	7.8	3.2	2.7	1.7	1.4	4.3
Diphthong	3.3	14.0	1.0	1.1	5.6	7.0	3.3	4.4	2.4	3.4	3.5

After statistical measures of significance (both parametric and non-parametric) were used, the following conclusions were made:

1. The position of mean duration from highest to lowest versus speaking style is shown below. Results from all speech classes using word list No. 1 are shown (* = significant with respect to neutral).

Shift in mean duration

Word 385 utterances (percent)			Vowel 429 phonemes (percent)			Consonant 759 phonemes (percent)			Semivowel 121 phonemes (percent)			Dipthong 88 phonemes (percent)		
SL	*	+73	SL	*	+84	C	*	+84	SL	*	+112	SL	*	+94
C	*	+39	A	*	+69	SL	*	+52	LM	*	+63	A	*	+64
A	*	+38	L	*	+58	SO		+24	A		+42	L	*	+53
L	*	+36	C		+26	C7		+22	Q		+40	LM		+30
LM	*	+20	LM		+24	C5		+12	C		+39	Q		+9
Q		+10	Q		+13	Q		+5	L		+27	SO		+9
SO		+7	N			LM		+4	C5		+20	C		+4
C7		+5	SO		-8	L		+3	SO		+19	N		
C5		+1	C5		-8	N			C7		+14	C7		-7
N			C7		-8	A		-12	N			C5		-8
F	*	-26	F	*	-28	F	*	-27	F		-27	F		-27

2. Mean word duration values may be used as significant indicators for speech in slow, clear, angry, loud, Lombard, or fast styles when compared to neutral conditions.

3. Slow and fast mean word duration are all significantly different from all styles considered .

4. Clear mean consonant duration was significantly different from all styles except slow.

5. Word and phoneme class duration are not significant indicators for moderate versus high task work load conditions.

Duration Variability

1. The position of duration variance from highest to lowest versus speaking style is shown below. Results from all speech classes using word list No. 1 are shown (* = significant with respect to neutral).

Shift in duration variance

Word 385 utterances (percent)			Vowel 429 phonemes (percent)			Consonant 759 phonemes (percent)			Semivowel 121 phonemes (percent)			Dipthong 88 phonemes (percent)		
SL	*	+173	A	*	+191	C	*	+456	A	*	+1045	SL	*	+324
A	*	+128	SL	*	+166	SL	*	+294	LM	*	+531	A		+112
C	*	+122	L	*	+141	L	*	+106	LM	*	+531	L		+70
L		+56	C	*	+115	A	*	+83	C	*	+370	Q		+33
LM		+32	LM		+65	C7	*	+78	L	*	+326	LM		+6
N			Q		+14	SO	*	+63	Q	*	+296	C7		+3
SO		-11	N			Q		+61	C5		+150	C		+1
Q		-11	C5		-4	LM		+44	C7		+106	N		
C5		-11	C7		-4	C5		+39	SO		+91	C5		-27
C7		-22	SO		-22	N			F		+45	SO		-66
F		-33	F	*	-54	F	*	-39	N			F		-70

2. Duration variance increased for all domains (word, vowel, consonant, semivowel, diphthong) under slow stress condition.

3. Duration variance decreased for most domains under fast stress condition.

4. Duration variance significantly increased for angry speech.

5. Duration variance generally increased for loud speech but was mixed for soft speech.

6. Clear consonant duration variance was significantly different from all styles.

7. Duration variance is not a significant indicator for moderate versus high task work load conditions.

Intensity

Measures of intensity and intensity variability were analyzed across the utterances in the data base. Tables 13 and 14 describe the results.

Table 13

Average Word and Speech Class Intensity Using Word List No. 1 (35 words/style, 11 styles = 385 words) Over Various Speaking Styles and Stress Conditions (RMS values)

Stress condition	Average duration (msec)										
	Neutral	Slow	Fast	Soft	Loud	Anger	Clear	Question	Cond50	Cond70	Lombard
Word	7663	7982	7812	7277	10561	11307	7067	8110	7075	6934	8286
Vowel	9610	9692	9404	9326	12002	12700	9786	10172	8857	8996	9699
Consonant	1394	1481	1425	1866	1164	1562	1287	1602	1592	1715	1401
Semi-vowel	10032	9323	9983	10072	9443	11629	8272	10043	8498	8353	8322
Diphthong	10125	9989	10460	9393	14800	14724	10394	10478	9807	9742	10913

Table 14

Variance of Average Word and Speech Class Intensity Using Word List No. 1 (35 words/style, 11 styles = 385 words) Over Various Speaking Styles and Stress Conditions (RMS variance x105)

Stress condition	Variance of average duration (msec)										
	Neutral	Slow	Fast	Soft	Loud	Anger	Clear	Question	Cond50	Cond70	Lombard
Word	12.3	8.5	10.1	16.0	31.2	50.7	6.5	17.1	9.2	8.3	16.8
Vowel	93.3	76.5	93.4	63.6	116.0	193.0	109.0	45.3	92.2	101.0	80.4
Consonant	21.8	32.1	20.1	35.8	26.0	33.6	24.9	32.9	24.1	33.1	23.9
Semi-vowel	128.0	106.0	136.0	102.0	231.0	571.0	75.7	109.0	162.0	187.0	152.0
Diphthong	38.7	14.6	29.0	22.1	17.3	19.5	23.6	11.5	23.2	30.8	8.4

After statistical measures of significance (both parametric and non-parametric) were used, the following conclusions were made:

1. The position of mean RMS intensity from highest to lowest versus speaking style is shown below. Results from all speech classes are based on word list No. 1 (* = significant with respect to neutral).

Shift of average RMS intensity										
Word 385 utterances (percent)			Vowel 429 phonemes (percent)			Consonant 759 phonemes (percent)		Semivowel 121 phonemes (percent)		Dipthong 88 phonemes (percent)	
A	*	+48	A	*	+32	SO	+33	A	+16	L	* +46
L	*	+38	L	*	+25	C7	+23	Q	0	A	* +45
LM		+8	Q		+6	Q	+15	SO	0	LM	+8
Q		+6	C		+2	C5	+14	F	0	Q	+4
SL		+4	LM		+1	A	+12	N		C	+3
F		+2	SL		+1	SL	+3	L	-6	F	+3
N			N			F	+2	SL	-7	N	
SO		-5	F		-2	LM	+1	C5	-15	SL	-1
C		-8	SO		-3	N		C7	-17	C5	-3
C5		-8	C7		-6	C	-8	LM	-17	C7	-4
C7	*	-10	C5		-8	L	-17	C	-18	SO	-7

2. Average RMS word intensity values may be used as significant indicators for speech in angry, loud, and high work load task styles when compared to neutral conditions.

3. Loud and angry average RMS word intensity are significantly different from all other styles considered.

4. Loud and angry average RMS vowel and diphthong intensities were significantly different from all other styles considered.

5. Average RMS consonant and semivowel intensity are not significant stress indicators for any of the styles considered.

6. Average RMS intensity is not a significant indicator for moderate versus high task work load conditions.

Intensity Variability

1. The position of average RMS intensity variance from highest to lowest versus speaking style is shown below. Results from all speech classes are based on word list No. 1 (* = significant with respect to neutral).

Shift in the variance of average RMS intensity									
Word 385 utterances (percent)			Vowel 429 phonemes (percent)			Consonant 759 phonemes (percent)			Semivowel 121 phonemes (percent)			Dipthong 88 phonemes (percent)	
A	*	+312	A	*	+107	SO	*	+64	A	*	+346	N	
L	*	+154	L		+25	A		+54	L		+80	C7	-20
Q		+39	C		+17	C7		+52	C7		+46	F	-25
LM		+37	C7		+8	Q		+51	C5		+27	C	-39
SO		+30	F		0	SL		+47	LM		+19	C5	-40
N			N			L		+19	F		+6	SO	-43
F		-18	C5		-1	C		+14	N			A	-50
C5		-25	LM		-14	C5		+11	Q		-15	L	-55
SL		-31	SL		-18	LM		+10	SL		-17	SL	-62
C7		-33	SO		-32	N			SO		-20	Q	-70
C		-47	Q		-52	F		-8	C		-41	LM	-78

2. Variance of average RMS word intensity values may be used as significant indicators for speech in angry and loud styles when compared to neutral conditions.

3. Variance of loud and angry average RMS word intensity is significantly different from most styles considered.

4. Variance of angry average RMS vowel and semivowel intensities were significantly different from most styles considered (15/20).

5. The variance of average RMS consonant and diphthong intensity were not significant stress indicators for most styles considered.

6. The variance of average RIS intensity (for word or phoneme class) was not a significant indicator for moderate versus high task work load conditions.

Table 15

Average Formant Frequencies for Phoneme /IY/

Stress condition	Number frames	Average formant frequencies phoneme: IY			
		F1	F2	F3	F4
Neutral	117	411	1970	2607	3368
Slow	228	393	2000	2660	3375
Fast	80	415	2021	2582	3385
Soft	102	404	1955	2617	3444
Loud	236	431	2071	2686	3414
Anger	247	586	2078	2661	3357
Clear	149	387	2086	2667	3379
Question	167	428	2064	2615	3489
Cond50	121	413	2042	2642	3360
Cond70	118	421	2044	2622	3367
Lombard	146	412	2006	2644	3376

Formants

Measures of formant frequencies, their variabilities, formant bandwidths, and their variabilities were analyzed across the utterances in the data base. Tables 15, 16, 17, and 18 describe the results for the vowel /i/ (as in beet).

Based on results from /i/, formant location and bandwidth appear to be reasonable indicators of stress. However, this discriminating ability may not hold for other phonemes. Therefore, an evaluation for another five phonemes was also performed. The results were quite successful. Of the 400 student t-tests, 166 were statistically different from neutral. most of these involved loud, angry, or Lombard formant information. In addition, a majority of the significant comparisons involved formant locations and bandwidths for F1 and F2. Of all the stress conditions, average formant information for loud and angry were the most consistent across the six phonemes.

The next area of interest is to determine if variability in formant location or bandwidth are valid stress indicators. As seen in Table 17, the variance of formant location showed large shifts in only limited cases. Formant bandwidth values from Table 18, however, showed much higher levels of change. Also, caution should be exercised in the use of variance as a stress relayer for F3, since corresponding pole-pairs may not represent true resonance. Statistical tests showed that variance of formant location was significantly different from neutral for exactly half of the styles and formants considered, with higher discriminating ability for F3 and F4. An interesting point can also be made between mean and variance of formant location. Half of the cases of significantly different means coincided with cases in which the variance was also significantly different. This indicates that styles that vary formant location, will also increase formant variability in conveying that stress condition. This was even more pronounced for formant bandwidths. Of the 28 significantly different bandwidth variance comparisons (with neutral), 24 coincided with mean bandwidth. Thus, if a talker varies the mean of a formant bandwidth, there is a high degree of certainty that bandwidth variability will also increase. Formant bandwidth variance for F1 and F2 showed a high degree discriminating ability (15/20). Significance in bandwidth variance was high for loud, angry, clear, questions, high task condition, and Lombard effect styles. The results show that variance of formant location and bandwidth is fair to good in differentiating stress parameters (47/80); pairwise comparisons are statistically different. This was also true for the other vowels considered. Bandwidth variance for Lombard, loud, and angry styles was predominantly different from neutral. Three other phonemes (/oU/, /N/, /R/) also showed significant variations in the variance of formant location and bandwidth. Similarity between styles that possessed significantly different means and variances from neutral was also demonstrated. Of a possible 71 coinciding significant comparisons of the mean and variance of formant location, 45 overlapped. For formant bandwidth, 58 of 81 significant comparisons overlapped. Thus, the previous findings of a high degree of correlation between variation in formant mean and formant variance for both location and bandwidth were supported.

Table 16

Average Formant Bandwidths for Phoneme /IY/

Stress condition	Number frames	Average formant bandwidths phoneme: IY B1	B2	B3	B4
Neutral	117	52	222	496	366
Slow	228	81	298	451	358
Fast	80	102	358	498	524
Soft	102	180	327	483	518
Loud	236	86	174	355	219
Anger	247	102	166	464	392
Clear	149	105	356	613	531
Question	167	62	423	436	855
Cond50	121	110	193	270	291
Cond70	118	111	182	218	296
Lombard	146	73	139	250	185

Table 17

Variance of Formant Frequencies for Phoneme /Y/

Stress condition	Number frames	Variance of formant frequencies phoneme: IY F1	F2	F3	F4
Neutral	117	1149	10822	38876	6814
Slow	228	827	9039	13383	7621
Fast	80	1168	11288	24476	25655
Soft	102	1930	10779	17860	8103
Loud	236	5267	10032	12233	4249
Anger	247	2985	15425	13460	10926
Clear	149	2108	20012	28449	27657
Question	167	15548	18264	33559	58478
Cond50	121	1804	16260	13284	8508
Cond70	118	1149	9676	16252	9004
Lombard	146	980	9902	11621	4207

Table 18

Variance of Formant Bandwidths for Phoneme /IY/

Stress condition	Number frames	Variance of formant bandwidths phoneme: IY B1	B2	B3	B4
Neutral	117	1001	30067	89893	64002
Slow	228	3068	38426	84537	41609
Fast	80	7699	56580	104689	179823
Soft	102	29146	28356	77665	134411
Loud	236	1639	36476	23483	9599
Anger	247	9657	16425	42033	58077
Clear	149	9379	155603	178626	339158
Question	167	2963	171664	114890	279899
Cond50	121	11280	22796	47982	58466
Cond70	118	7583	9889	13027	36425
Lombard	146	1474	8172	15544	5180

For a full discussion of the statistical tests and the results (more than 200 pages), refer to Hansen (1988).

Glottal measurements of Speech in Stress

In this portion of the report, the glottal wave form measurement techniques and the resulting analyses are described.

Accurate modelling of the glottal source has long been a subject of great interest in speech processing. There are many applications for glottal modelling including speech coding (Bergstrom & Hedelin, 1989), synthesis (Carlson, Fant, Gobl, Karlsson, & Lin, 1989), and recognition (Hansen & Clements, 1989a). The goal of this research has been to describe the characteristics of the glottal wave forms of 11 types of stressed speech and to use this knowledge to improve automatic recognition of stressed speech. This improvement could be made either by directly incorporating glottal effects into the recognition process or by identifying the speech style with its glottal wave form and choosing an appropriate recognition-compensation algorithm.

It is generally believed that one of the major conveyors of stress is the manner of glottal excitation (i.e., the glottal source wave form). If one wishes to study how stress and speaking style affect speech, it is important to be able to examine the glottal source wave form. Unfortunately, most existing analysis methods do not allow for convenient separation of the glottal and vocal tract effects; hence, reliable information about changes in the glottal wave form caused by stress is difficult to obtain. Additionally, extracting the glottal wave forms from stressed speech presents specific difficulties. The most significant difficulty derives from the fact that the pitch period and other characteristics vary across speech styles. For example, in angry and loud speech, both the pitch period and the interval of glottal closure are very short. Further, question exhibits a pitch period that rapidly changes during an utterance. The vocal tract can also be very difficult to model under certain stress conditions such as 50 tasking and 70

tasking. Because it is important to maintain consistency in the method of extraction over the styles of speech, all the differences in the stressed speech must be accounted for in the design of the extraction method.

In this report, a method for extracting the glottal wave form from a segment of voiced speech is presented. This method is based on a procedure suggested by Wong, Markel, and Gray (1979) which has been specifically tailored for this application. The glottal wave forms for the 11 types of stressed speech contained in the MIT Lincoln Labs multi-style speech data base (normal, angry, loud, soft, slow, fast, clear, question, 50% tasking, 70% tasking, and Lombard) have been extracted from utterances of the vowel /I/ in fix" and "six" and the vowel /ɛ/ in "destination." These glottal wave forms have been analyzed statistically and qualitatively to identify glottal characteristics that are unique to a given stress style.

Theory

In the standard speech model, we assume

$$S(z) = G(z)V(z)R(z), \tag{58}$$

in which

$S(z)$ = z-Transform of speech wave form, $s(n)$

$G(z)$ = z-Transform of glottal source wave form, $g(n)$

Therefore,

$V(z)$ = vocal tract filter

$R(z)$ = radiation impedance at the lips

$$G(z) = \frac{S(z)}{V(z)R(z)} \tag{59}$$

Thus, $g(n)$ can be obtained by inverse filtering $s(n)$ by $v(n)$ and $r(n)$. For the voiced non-nasalized phonemes used in this research, the vocal tract can be modelled as an all-pole filter.

Radiation impedance at the lips has often been modelled as a filter consisting of one or two zeroes. Based on the work of Barnwell, Schafer, and Bush (1977), this radiation is modelled as a zero pair and a pole pair. In Barnwell's research, it was hypothesized that one difference between LPC synthesized speech and natural speech is a result of the glottal pulse being non-minimum phase. This can be compensated by using a fixed second order pre-emphasis filter (i.e., two zeroes) and a 10-pole vocal tract model (for 8-kHz sampled speech). These results were incorporated by modelling the effects of the radiation at the lips impedance as a pole pair and a zero pair.

Covariance LPC analysis minimizes the error, $e(n)$, in which

$$e(n) = s(n) + \Sigma_k a(k)s(n - k) \tag{60}$$

If the coefficients, $a(k)$, model the vocal tract perfectly and the radiation impedance at the lips has been compensated, $e(n)$ would equal the glottal source, $g(n)$. The estimation of the vocal tract parameters is most accurate during the period of glottal closure since at this time, $s(n)$ is theoretically a freely decaying oscillation affected only by the vocal tract and radiation at the lips. Therefore, an all-pole model for $s(n)$ is a good assumption during glottal closure and produces a relatively accurate model of $v(n)$. In

this procedure, covariance LPC analysis is performed over windows no longer than the expected period of glottal closure of s(n), shifted by one sample at a time. When the error wave form corresponding to this analysis is examined, glottal closure is assumed to exist over the segment for which e(n) is close to zero. Since the vocal tract has constraints on how rapidly it can change, the model found during the identified glottal closure is used to inverse filter an interval long enough to include four pitch periods of s(n)--the output being four periods of g(n).

Method

To compare and draw conclusions about changes in the glottal wave form caused by stress and speaking style, it is important that the extractions be performed under the same conditions in each case. The actual modelling of the vocal tract must be done over the same portion of a glottal period to minimize differences between the glottal wave forms of different speaking styles caused by differences in the modelling technique. This poses a particular problem because of the variability of the pitch period for different types of stressed speech. Such styles as angry and loud have pitch periods which are so short that the vocal tract cannot be accurately modelled with the samples comprising glottal closure. There is, therefore, a tradeoff between using the ideal segment (glottal closure) and maintaining consistency across the speaking styles by using the same length segment of a glottal period for each extraction. To maximize the number of samples in the vocal tract model, to maintain the accuracy of the model, and to maintain consistency across the stress styles, the extractions were performed over segments slightly longer than the apparent glottal closure.

The first step in the extraction of the glottal source wave form from a given utterance is computing e(n) using covariance LPC analysis. The LPC analysis is performed over a segment of windowed speech using a 10-pole filter (four pole pairs for the vocal tract and one pole pair for radiation at the lips). This window is shifted by one sample at a time, and e(n) is computed for each window of modelled speech. The error wave form is then examined and a starting point is chosen where e(n) is close to zero for a number of samples. The vocal tract is again modelled at this point, and this model is used to inverse filter four pitch periods of the speech wave form, s(n). This result is integrated twice to account for the previously described zero pair in radiation at the lips to produce the glottal source wave form, g(n). The analysis window is then shifted by one sample, the vocal tract is modelled again, and the inverse filtering and integration are repeated to produce a new g(n). This is done five to ten times and the best g(n) is chosen. In general, the extracted waveforms are extremely similar in shape and form. These iterations are conducted to find the best g(n) for those stress styles which are more difficult to model (e.g., 50% tasking, 70% tasking, and Lombard). After careful study, a period of a glottal wave form is described by six parameters: opening slope, closing slope, opening duration, top duration, closing duration, and closed duration. These parameters are illustrated in Figure 6. The four duration parameters are measured as the number of samples between the endpoints and including one endpoint. The two slope parameters are measured as the slope at the onset of opening and the offset of closing.

At this point, the amplitude of g(n) depends only on the error, e(n), at the point where the inverse filter was computed. This error can be affected by many things. To make accurate comparisons of the areas and slopes of the various styles of glottal wave forms, all the amplitudes were normalized with respect to normal. First, all glottal wave forms are

normalized to have the same maximum amplitude. Each wave form is then multiplied by a factor of (given stress intensity)/(normal intensity) and divided by (given stress open-closed ratio)/(normal open-closed ratio). The intensity figures used are from previous research (Hansen, 1988) and the open-closed ratios (ratio of open part of glottal pulse to closed part of glottal pulse) used are from this research. In this way, the amplitudes compared to normal are meaningful and comparisons between the styles are valid.

This process was performed on six to eight utterances for each of the 11 stress styles resulting in 50 to 100 pitch periods of glottal source for each style of speech. Both the /I/ and /ɛ/ contexts were used to ensure that the results were not vowel-specific. Furthermore, the glottal wave forms from a second speaker were examined to ensure that the results were not speaker-dependent. Each pitch period of glottal source was hand-marked to extract the six parameters, producing 66 distributions--one for every stress style for each parameter.

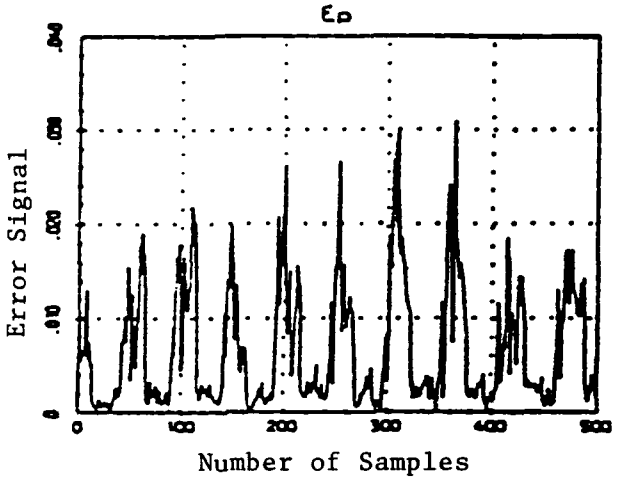


Figure 6. Error wave form for normal speech.

Several statistical tests were performed on these distributions. The Kolmogorov Smirnoff (K-S) test, which is a non-parametric test that compares two distributions to decide if the distributions are not significantly different, was used in two ways. First, the K-S test was performed on each pairwise comparison of the 66 distributions. Secondly, the K-S test was used to compare each of the 66 distributions to its respective best fit gaussian distribution. This was done to ascertain whether the assumption that the distributions were gaussian was a fair assumption. To confirm the results of the second set of K-S tests, a chi-square goodness-of-fit test was also performed, comparing each of the 66 distributions to its best fit gaussian. On the basis of the results of these two tests, the best fit gaussian mean and variance were computed for each of the 66 distributions.

Results

Example glottal wave forms for each of the 11 speaking styles are presented in Figure 7. The K-S test versus a gaussian and the chi-square goodness-of-fit test for a gaussian both showed that the assumption that these distributions are gaussian is not a bad one. The best fit gaussian means are provided in Table 19. The most interesting results obtained from examining the statistics in Table 19 are presented in Table 20. An important result is that, except for *question*, the percentage of the pitch period during

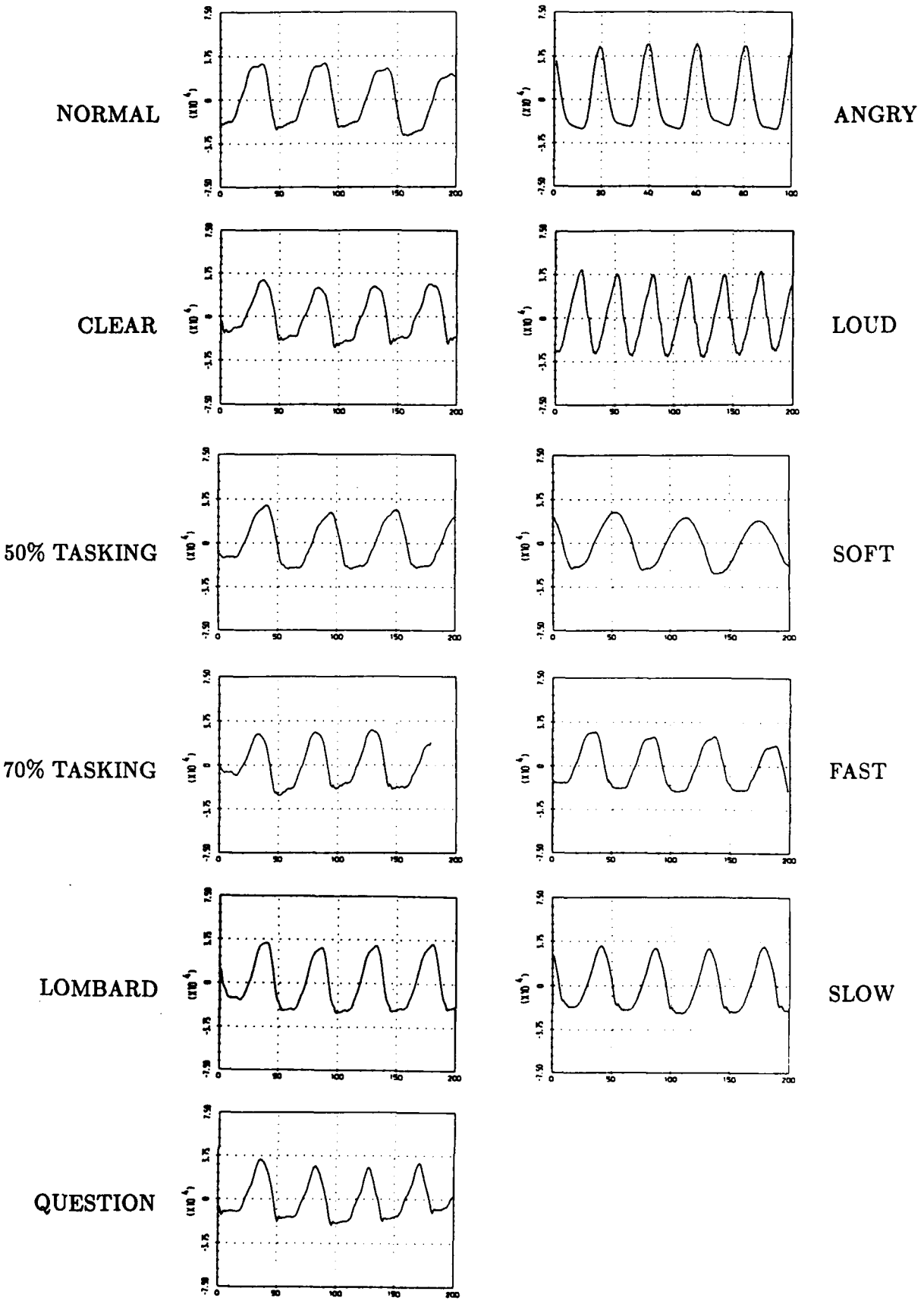


Figure 7. Example extracted glottal wave forms from the 11 types of stressed speech. (All horizontal axes are in samples; speech has been sampled at 8 kHz. All amplitudes are normalized.)

Table 19

Best Fit Gaussian Means

	Closing slope	Opening slope	Closed duration	Closing duration	Opening duration	Top duration
Normal	-4798	2643	17.7	10.2	15.6	9.9
Slow	-4786	2692	16.9	10.2	15.5	8.7
Fast	-392	2376	15.5	11.0	16.0	8.4
Soft	-2632	1921	17.7	14.7	18.6	9.9
Loud	-9298	3532	6.3	6.9	17.0	2.9
Angry	-9910	9198	9.1	6.3	6.9	2.0
Clear	-5011	2686	15.8	9.5	16.0	6.9
Question	-4831	3034	14.0	9.4	14.9	7.0
50%	-4522	2321	17.3	11.1	16.0	9.8
70%	-4100	2138	16.7	10.7	15.7	9.9
Lombard	-5430	2871	15.2	9.3	15.2	7.6

Table 20

Results Generated Using Best Fit Gaussian Means

	Closing versus opening slope	Closing versus opening duration	Percent pitch period closed
Normal	1.8	.65	31
Slow	1.8	.66	30
Fast	1.7	.69	29
Soft	1.4	.79	30
Loud	2.6	.40	23
Angry	1.1	.92	31
Clear	1.9	.60	30
Question	1.6	.63	36
50%	2.0	.69	30
70%	1.9	.68	30
Lombard	1.9	.60	31

which the glottis is closed is fairly constant across the speaking styles. More significant are the variations across the speaking styles of the ratios of closing to opening slope and closing to opening duration. These ratios are different enough that they may be used to identify a speaking style. Considering the ratio of closing to opening slope, angry, fast, loud, question, and soft are significantly different from each of the other speaking styles. Similarly, angry, loud, and soft have quite different closing to opening duration ratios.

The results from the pairwise K-S tests are important for the task of identifying a speaking style with its glottal wave form if a Bayesian classifier is used. For this to work, the distributions must be

significantly different for each speaking style using a linear combination of the six parameters. The results from the K-S pairwise tests show that the parameters that contain the most information about a speaking style are closing slope, opening slope, and closed duration. This is as expected since the slopes of glottal opening and closing provide most of the control over the amount of acoustic energy produced. Furthermore, since the duration of glottal closure is approximately a constant percentage of the pitch period for a given speaking style, the closed duration parameter directly reflects the length of the pitch period. It is well accepted that one important conveyor of stress is the length of the pitch period. Further examination of statistics from the tests indicate that each speaking style is significantly different from each of the other ten using combinations of the six parameters.

Qualitative Observations

Normal

The glottal wave shape in normally spoken speech is characterized by a slope of opening that is slower than the slope of closing. The pulse has a distinct closed period that is about 1/3 of a pitch period. The transition from opening to closing is slow (i.e., the pulse has a flat top).

Clear

The glottal wave shape of clear speech is not drastically different from the normal wave shape. In the case of clear, the transitions from closing to closed and from closed to opening are more abrupt than in normal. As in the case of normal, clear exhibits a period of distinct closure and an opening slope that is slower than the slope of closing. The top of the pulse is more "peaky" in clear.

50% Tasking

Speech that has been produced under 50 tasking conditions also is not extremely different from normal speech. It is, however, much more difficult to model the vocal tract in this case. The glottal wave shape exhibits a slower slope of opening than of closing, although this difference is not as distinct as it is in the case of normal. The pulse, like normal, has a distinct closed period and smooth transitions from closing to closed and from closed to opening. The transition from opening to closing is smooth but not as fat as in normal speech.

70% Tasking

The glottal wave shape in the case of 70% tasking is very much like that in 50% tasking. The major difference, which is that the width of the pulse is smaller in the 70% tasking, is mainly because the pitch period is smaller in the case of 70% tasking.

Lombard

Lombard speech, which is produced when the speaker is in a noisy environment, is more like clear than like loud. This implies that the speaker does not merely try to speak more loudly than the noise; he tries to make the speech distinct and clear. The glottal wave shape in the case of Lombard speech has a distinct closed phase, slower opening than closing, and a fairly "peaky" top.

Angry

The angry glottal wave shape, as is expected, is significantly different from the normal wave shape. Angry is the condition in which the glottis is the most completely closed. The pulse shape is very "peaky," marking an abrupt change from opening to closing. The amplitude of the pulses is very high. The slopes of opening and closing are steep and are not significantly different. The transitions from closing to closed and from closed to opening are abrupt. This is expected since the acoustic energy generated is maximized with extremely sharp glottal closure.

Loud

The loud glottal wave shape is also very different from normal. It is not, however, much like the angry wave shape except in that the amplitude is very high. Loud glottal wave shapes are characterized by a short period of glottal closure, very steep slopes, and extremely abrupt transitions. Again, glottal closure occurs abruptly to maximize the acoustic energy produced.

Soft

Soft speech, which approaches breathy or whispering, speech, has a glottal wave shape that is characterized by very smooth transitions and slow opening and closing. The wave form is almost sinusoidal in nature. The glottis does not fully close. This condition is the opposite of loud. Here, less acoustic energy is generated because of the slow and incomplete glottal closure.

Fast

The fast glottal wave shape opens at about the same rate as it closes. The top of the pulse is flat. Glottal closure is distinct and flat. This would indicate an effort to make the speech clear and to generate acoustic energy efficiently.

Slow

The glottal wave shape in the case of slow speech is most significantly different from that in fast speech in that the transitions to and from glottal closure are much less distinct and more smooth. The flattening at the top is smaller, making the overall pulse shape almost continuously smooth.

Question

The most significant characteristics of the glottal wave shape in the case of question occur over the length of the vowel (20+ pitch periods). The amplitude of the pulses and the pitch period are constantly decreasing. These two facts are most likely the reason that the pulse shape in question deviates from normal. The two types of wave shape differ in that the slopes of opening and closing in question are not significantly different. Further, in the case of question, the pulse shape tends to be more peaky.

Conclusion

This method of extracting glottal wave forms using inverse filtering, while time-consuming, gives reasonable results. Because the extraction process is extremely sensitive to the point at which the modelling is begun and the length of data over which the speech is modelled, most of this process must be interactive.

The extracted wave forms are consistent across utterances, and changes in the glottal wave shape that theoretically should occur under different stress conditions are present in the extracted glottal wave forms. Although not enough utterances were employed to compute precise statistical descriptions of the wave forms, this method, if used on a sufficient number of wave forms, would establish statistically significant trends. These results would make it possible to parameterize the glottal wave form for application to robust automatic speech recognition.

The results of this research show that each of the 11 styles of glottal wave forms has a unique profile based on the six parameters: closing slope, opening slope, opening duration, top duration, closing duration, and closed duration. This is not a profile that depends on the vowel spoken. Furthermore, examination of the glottal wave forms of a second speaker suggests that the results are not speaker-dependent. Once the glottal extraction procedure has been automated, these statistical profiles can be used to identify the style of a given utterance of speech from a voiced segment of that utterance. Once the speaking style is known, automatic recognition of the speech can be improved by choosing a specifically designed enhancement algorithm or a specifically trained codebook. These results should allow for significant improvement of automatic recognition of stressed and otherwise style-variant speech. Another possibly useful application would be that of changing one speaking style into another by LPC synthesis using appropriately modified excitation wave forms.

Stress Compensation Algorithms

The motivation for the analysis of speech under stress was to uncover those acoustically correlated parameters that vary during stressful conditions. Variation in these parameters may suggest a possible explanation for adverse recognition performance in diverse environments. The previous investigation explored areas of speech production which traditionally have not been associated with present day recognition algorithms. The reasons for this are twofold. First, it may be possible to improve existing recognition algorithms by allowing preprocessors to reduce or eliminate parameters that are affected by stress. Although the effects of some parameters (i.e., pitch, duration, intensity) are somewhat mitigated in many recognition procedures, severe variations do adversely affect recognition performance (e.g., HMM recognizers have certain "time constants" which tolerate only a limited degree of duration variability). Other stress analysis parameters such as characteristics of glottal source spectrum (spectral tilt, energy distribution) and vocal tract (formant center frequencies, bandwidths, spectral tilt, and the variability of these) have direct consequences in recognition performance. Second, other stress-relaying parameters not used for recognition may be used to reliably identify when an utterance is stressed so that appropriate stress compensation can be employed.

A set of stress compensation algorithms was formulated based on results from three stress analysis domains. These approaches assume the stress condition (e.g., loud, angry, clear, etc.) to have already been identified. The algorithms are based on obtaining a table of compensation factors for all phonemes, for each stress condition. The three possible processing steps include compensating for (a) average formant location (F1,F2,F3,F4), (b) average formant bandwidth (B1,B2,B3,B1), and (c) overall word intensity. To calculate formant location and bandwidth values, root solving and pole ordering of the LPC polynomial for each speech frame was performed. To reduce the variance of average formant location and bandwidth estimates, a smoothing operation was performed before average formant values were calculated. This served to improve the estimation of average values by reducing the effects of outlying values caused by misclassification during ordering. Formant compensation factors were obtained by taking the ratio of average formant values between neutral and stressed conditions. Parametric and non-parametric statistical tests were used to verify significance of variation in formant characteristics. Table 21 presents sample compensation factors for average formant location and bandwidth used for the angry stressed condition. As an example, consider the compensation for the vowel /e/. The term F1 (= 0.63) was used to decrease all first formant locations for phoneme /e/ under angry conditions. The average first formant location will then have the same average value as that found in neutral conditions. This process was repeated for each formant location and bandwidth. A similar table was obtained for each of the ten stress conditions. Once the compensation tables are known, preprocessing stress compensation algorithms were implemented. Two additional points are necessary. First, unlike the fully automated constrained speech enhancement algorithms summarized previously, the stress compensation algorithms require both knowledge of the type of stress and phoneme boundaries in order to apply compensation factors. Further research is underway to incorporate general stress compensation within the constrained enhancement algorithms, thereby removing this requirement. Second, as demonstrated in Hansen and Clements (1987), such compensation schemes could be easily be extended to LSP parameters, and therefore integrated within the speech enhancement algorithms. This particular approach was chosen since computational requirements were not an issue, and shifts in formant location and bandwidth gave a more intuitive feel for how the vocal tract spectrum was being adjusted.

Recognition Framework and Results

Advances made in the analysis of speech under stress and speech enhancement domains are joined to address the final goal of recognition in noisy stressful environments.

A fairly standard, isolated word, discrete observation HMM recognition system was used for evaluation. This system was LPC-based and had no embellishments. In all experiments, a five-state, left-to-right model was used. The system dictionary consisted of 20 highly confusable words from the second and third domains of the speech under stress data base. These words are also used by Texas Instruments and MIT Lincoln Labs to evaluate recognition systems. Subsets include {go, oh, no}, {six, fix}, and {wide, white}. Thirty-two examples of each word were used in the evaluation, six neutral examples for recognition, and two examples for each of the ten stressed speaking styles (i.e., soft, loud, etc.) for recognition (i.e., all tests fully open employing neutral training data). The 20 models employed by the HMM recognizer were trained using the forward-backward algorithm.

Figure 8 illustrates the recognition scenarios in the evaluation. Results from each are summarized in Figure 9. The first four evaluations establish base line recognition scores for neutral, stressful, noisy neutral, and noisy stressful speech conditions. The recognition rate of noise free neutral speech (88) confirms the confusability of the chosen vocabulary. Independent evaluations of this system with distinct vocabularies resulted in recognition rates of 100 (Hansen, 1988). Base line scores indicate that stress, with and without background noise, has a profound effect on recognition performance. Recognition rates dropped by an average 31% for stressful speech, with an additional 19% for noisy stressful speech, thus indicating that recognition degrades rapidly whether a speaker is undergoing stress, in noise, or a combination.

The fifth recognition scenario employed enhancement pre-processing of noisy neutral speech. In a study by Hansen and Clements (1988), the constrained enhancement algorithms were shown to be superior to implementations of past enhancement techniques (e.g., spectral subtraction, noncausal Wiener filtering) in preprocessing for recognition of noisy neutral speech. Therefore, only the constrained enhancement techniques are considered here. The constrained enhancement algorithm used (FF-LSP:T,Auto:I) was based on fixed frame constraints applied across time, and constraints applied to auto-correlation lags across iterations (see Hansen & Clements, 1988, and Hansen, 1988, for further discussion). The noise degradation was additive white gaussian, with SNRs determined over entire utterances. A 34% increase in recognition was observed for enhanced neutral speech. For the sixth recognition scenario, the same enhancement pre-processing was employed for noisy, stressful speech. Recognition rates significantly increased for all types of stress (an average +17.8%). It should also be noted that SNRs in low energy non-sonorant portions which discriminate confusable pairs (e.g., go - oh - no) may well be 20 dB lower than global SNR measurements. The enhancement preprocessors are therefore successful in reducing background noise as well as reducing some vocal tract variations caused by stress.

Next, stress compensation pre-processing of noise-free stressful speech was considered. Three stress compensation algorithms were evaluated: (a) average formant location compensation (FL), (b) average formant bandwidth compensation (FB), and (c) combined formant location and bandwidth compensation (FL+FB). All compensators included intensity compensation. Figure 9, presents results from these evaluations. Collectively, nine of the ten stressed conditions benefited from stress compensation. FL+FB is preferable for varying vocal effort (soft,loud) and angry speech (half of all recognition errors were eliminated). Stress compensation did not improve recognition performance for the clear speaking style, thereby suggesting that other stress factors (beside formant location and bandwidth) should be considered. Finally, for speech under the Lombard effect, FB compensation provided the best recognition improvement (+13%). Overall recognition performance was consistent across varying stress styles, indicating the success in reducing effects caused by stress.

The final recognition evaluation combined enhancement and stress compensation pre-processing. In half of the noisy stressful conditions, compensation did not appreciably raise recognition rates over enhancement pre-processing alone, thus suggesting that either enhancement pre-processing has the preformed necessary stress compensation, or that other forms of compensation are required. Improvement was observed in several key stress styles (e.g., loud, angry, Lombard). Increased recognition ranged from +22% to +27% over enhancement pre-processing alone, and +35% to +43% over original noisy stressful speech.

Table 21

Compensation Factors for Average Formant Location (F1,F2,F3,F4) and Bandwidth (B1,B2,B3,B) of 25 Phonemes for Angry Speech

Category	Phoneme	FORMANT COMPENSATION FACTORS F1	F2	F3	F4	B1	B2	B3	B4
Consonants									
Nasals	/N/	0.74	1.04	0.99	1.01	0.71	1.32	1.41	1.39
Stops									
Voiced	/D/	0.92	0.96	1.02	0.98	1.26	0.81	0.65	0.93
	/G/	1.11	0.80	0.96	0.98	1.70	0.73	1.05	1.10
Unvoiced	/T/	0.81	0.94	1.02	1.00	0.78	0.93	0.90	0.95
	/K/	0.96	0.99	1.02	1.02	0.84	0.53	0.87	0.60
Whisper	/H/	0.85	0.95	0.95	1.04	0.97	0.66	1.25	0.67
Affricates	/TSH/	1.27	0.93	1.02	1.01	0.96	1.06	1.02	1.08
Fricatives									
Voiced	/TH/	1.24	1.05	1.05	1.02	1.03	1.10	0.76	1.01
	/Z/	0.63	0.84	1.03	1.02	0.89	1.47	1.45	1.14
	/ZH/	1.32	0.95	1.08	0.99	1.85	1.64	0.98	0.74
Unvoiced	/F/	0.63	0.78	0.85	0.98	0.67	0.83	1.18	0.65
	/THE/	0.96	0.91	1.00	1.00	0.89	1.39	0.63	1.28
	/S/	0.88	0.93	1.01	1.07	1.28	0.72	0.96	0.93
Vowels									
Front	/IY/	0.70	0.93	0.96	0.98	0.60	1.38	1.09	0.92
	/I/	0.69	0.97	0.96	1.00	0.73	0.89	0.96	0.89
	/E/	0.74	0.77	0.99	0.99	0.65	1.26	0.41	1.32
	/AE/	0.63	0.96	0.96	1.03	0.84	1.42	1.25	0.73
Mid	/ER/	0.77	0.61	0.95	0.98	0.87	0.78	0.59	0.95
	/e/	0.63	0.92	0.97	0.95	0.61	0.96	0.68	1.19
Diphthongs									
	/AI/	0.60	0.77	0.90	0.98	0.60	0.37	1.28	1.00
	/EI/	0.68	0.81	1.03	1.05	1.40	1.09	0.88	0.78
	/oU/	0.60	0.94	1.00	1.01	0.96	1.36	1.49	1.13
Semi-vowels									
Liquid	/W/	1.02	1.51	1.05	1.00	1.26	1.35	1.12	0.93
	/L/	0.65	0.90	0.92	0.97	0.36	0.31	0.76	0.35
Glide	/R/	0.71	0.83	0.82	0.96	1.25	2.09	0.99	0.78

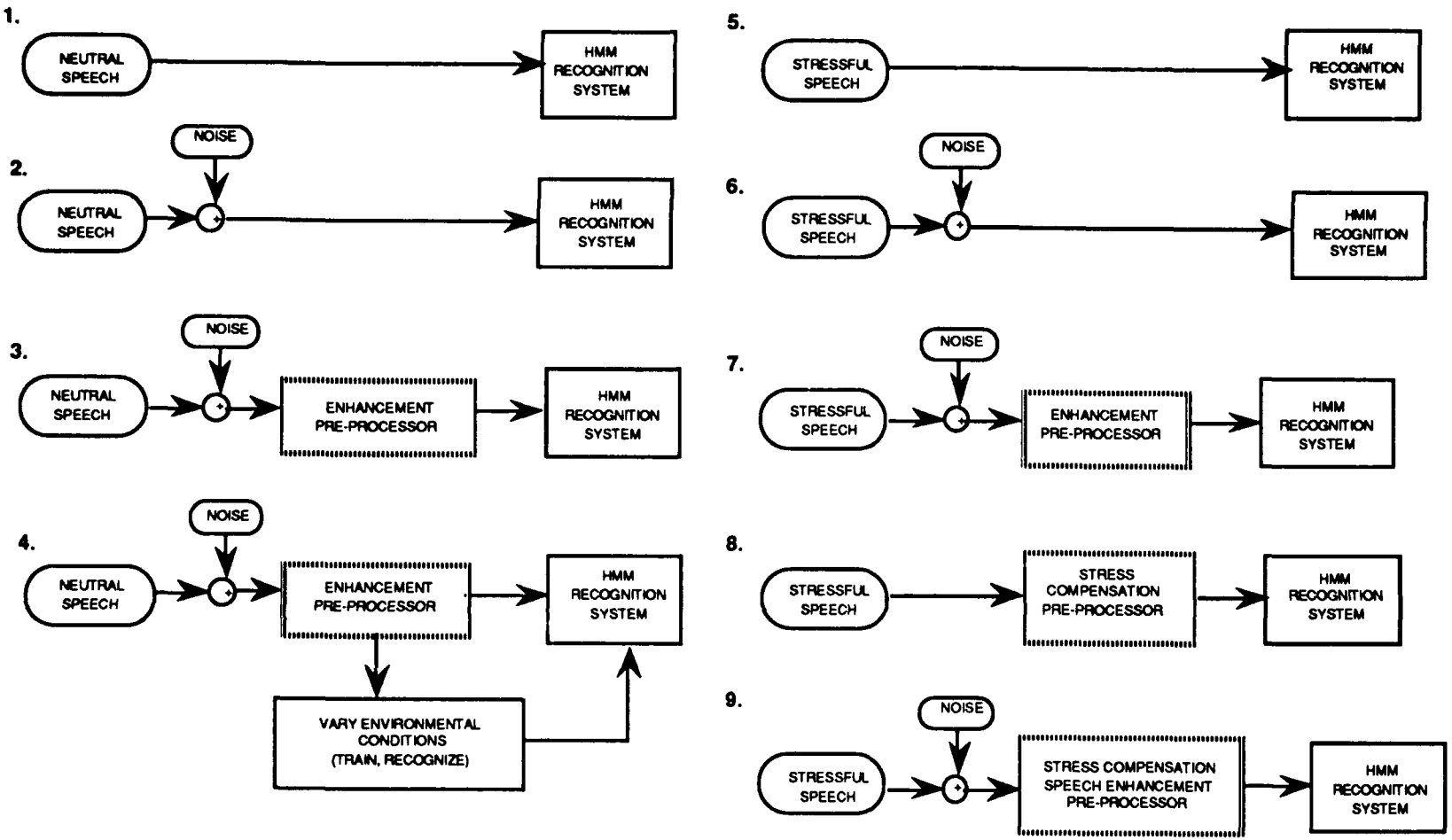
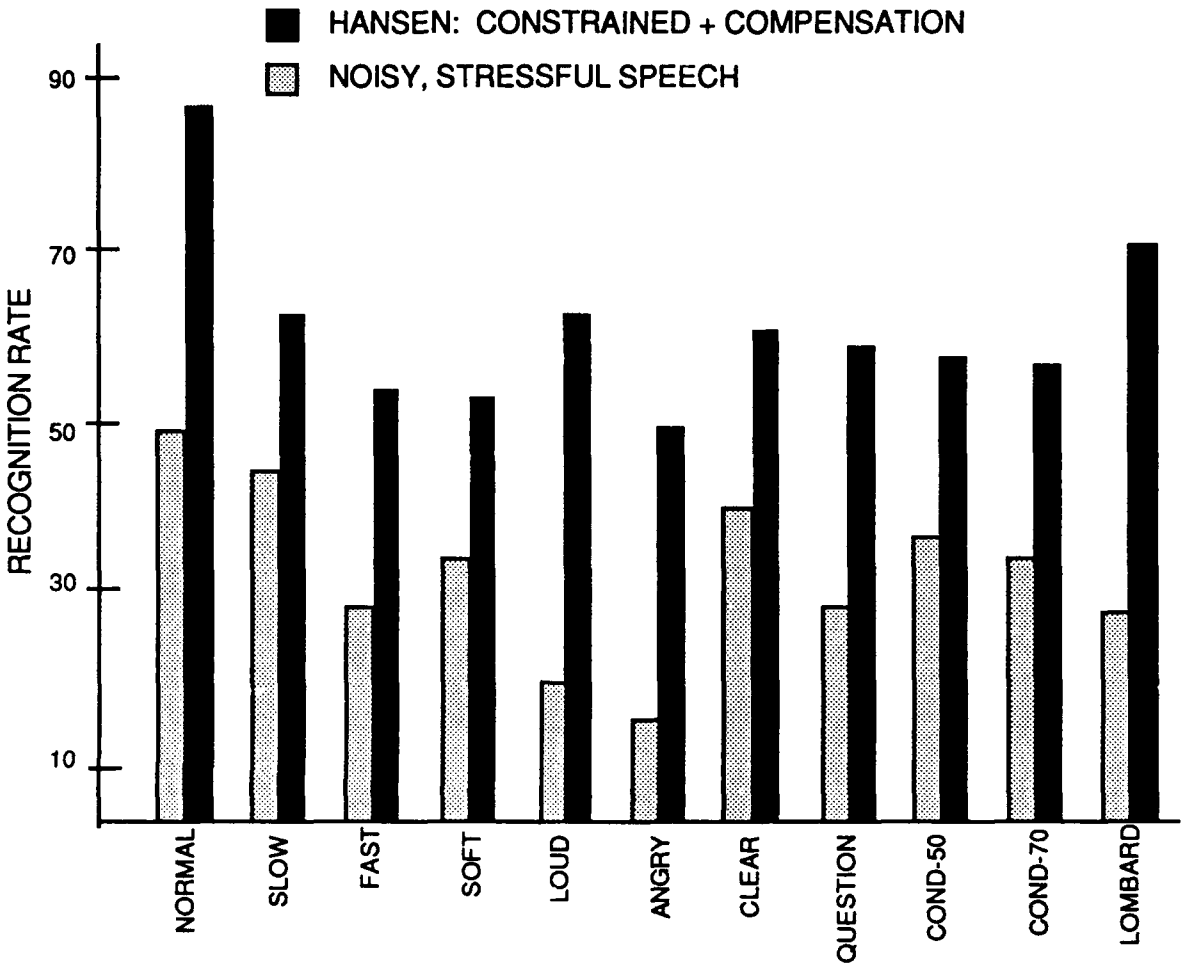


Figure 8. Robust automatic speech recognition scenarios. (The nine environments assume various levels and forms of stress and noise. Environments 1, 2, 5, and 6 establish base line recognition scores. The other five scenarios vary the type of stress, recognition procedure, and type of enhancement or stress compensation pre-processing.)



ENHANCEMENT & STRESS COMPENSATION RECOGNITION RESULTS IN PERCENT†

Condition	Neutral	Slow	Fast	Soft	Loud	Angry	Clear	Question	C50	C70	Lombard
Stressful, noisy	49	45	28	33	18	15	40	28	35	33	28
FF-LSP:T, Auto:I	83	57	53	43	35	28	58	55	58	53	38
plus compensator FL		50	47	53	50	45	47	35	50	55	55
plus compensator FB		56	37	53	56	25	58	56	55	55	70
plus compensator FL+FB		61	53	53	61	50	53	56	55	55	65

Figure 9. Recognition performance of noisy stressful speech, noisy stressful speech with constrained enhancement pre-processing (FF-LSP:T, Auto:I) and three combined speech enhancement-stress compensation algorithms: FF-LSP:T,Auto:I + FL,FF-LSP:T,Auto:I + FB,FF-LSP:T, Auto:I + FL+FB. (The bar graph illustrates the best recognition improvement possible employing stress compensation techniques. †Additive white gaussian noise, SNR = +30dB)

These results are encouraging since recognition in loud, angry, and Lombard conditions are most closely associated with speech from actual noisy stressful environments (such as an aircraft cockpit). The graph in Figure 9 summarizes the best combined enhancement, stress compensation pre-processing results. An inspection reveals consist recognition performance over varying noisy stress conditions, thereby indicating the effectiveness of pre-processing robust speech recognition.

Conclusions

The problem of speech recognition in noisy, stressful environments has been addressed in this report. A series of speech enhancement and stress compensation pre-processing algorithms was formulated that produce speech or recognition features which are less sensitive to varying factors caused by stress and noise. Previous results have shown the constrained enhancement algorithms to improve recognition performance for neutral speech over past enhancement techniques for a wide range of SNRs. Enhancement pre-processing also results in marked increases in recognition under noisy stressful conditions. Stress compensation techniques (based on formant location, bandwidth, and intensity) have been shown to reduce the effects of stress present in changing vocal tract characteristics, thereby improving recognition of noise-free stressful speech. Finally, combined stress compensation, speech enhancement pre-processing increased recognition rates by an average +27% (e.g., +43% loudly spoken speech, +42% speech spoken under the Lombard effect). In conclusion, combined speech enhancement and stress compensation pre-processing has been shown to be extremely effective in reducing the effects caused by stress and noise for robust automatic recognition.

REFERENCES

Barnwell, T. P., Schafer, R. W., & Bush, A. M. (1977). Tandem interconnections of LPC and CVSD digital speech coders (Technical Report E21-685-77-TB-2). Atlanta, GA: Georgia Institute of Technology.

Bergstrom, A., & Hedelin, P. (1989). Code book driven glottal pulse analysis. Proceedings of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing, S1, 53-56.

Boll, S. (1979). Suppression of acoustic noise in speech using spectral subtraction. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-27, 113-120.

Carlson, B., & Clements, M. A. (1990). A weighted projection measure for robust speech recognition. Proceedings of the IEEE Southeastcon.

Carlson, R., Fant, G., Gobl, C., Karlsson, I., & Lin, Q. (1989). Voice source rules for text-to-speech synthesis. Proceedings of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing, S1, 223-226.

Clements, M. A., & Lim, S. (1987). Hidden markov model speech recognition based on kalman filtering. Proceedings of the 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing.

Crosmer, J. R. (1986). Very low bit rate speech coding using the line spectrum pair transformation of the LPC coefficients (Doctoral dissertation, Georgia Institute of Technology, 1985). Dissertation Abstracts International, 46, 3954B.

Cummings, K. E., & Clements, M. A. (1990). Analysis of glottal waveforms across stress styles. Proceedings of the 1990 IEEE International Conference on Acoustics, Speech, and Signal Processing.

Cummings, K. E., Clements, M. A., & Hansen, J. H. L. (1989). Estimation and comparison of the glottal source waveform across stress styles using glottal inverse filtering. Proceedings of the IEEE Southeastcon, 776-781.

Folds, D., Gerth, J. I., & Engelman, W. R. (1986). Enhancement of human performance in manual target acquisition and tracking (Final Technical Report US-AFASM). Atlanta, GA: Georgia Institute of Technology, School of Psychology.

Hansen, J. H. L. (1989). Analysis and compensation of stressed and noisy speech with application to robust automatic recognition (Doctoral dissertation, Georgia Institute of Technology, 1988). Dissertation Abstracts International, 50, 293B.

Hansen, J. H. L., & Clements, M. A. (1985). Enhancement of Speech Degraded by Non-White Additive Noise, (Final Technical Report DSPL-85-6). Atlanta, GA: Georgia Institute of Technology.

Hansen, J. H. L., & Clements, M. A. (1987). Iterative speech enhancement with spectral constraints. Proceedings of the 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, 189-192.

Hansen, J. H. L., & Clements, M. A. (1988). Constrained iterative speech enhancement with application to automatic speech recognition. Proceedings of the 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing, 561-564.

Hansen, J. H. L., & Clements, M. A. (1989a). Stress compensation and noise reduction algorithms for robust speech recognition. Proceedings of the 1989 IEEE International Conference on Acoustics, Speech, and Signal Processing, 266-269.

Hansen, J. H. L., & Clements, M. A. (1989b). Use of objective speech quality measures in selecting effective spectral estimation techniques for speech enhancement. Proceedings of the Midwest Symposium on Circuits and Systems.

Hansen, J. H. L., & Clements, M. A. (1991). Constrained iterative speech enhancement. IEEE Transactions on Acoustics, Speech, and Signal Processing, 745-806.

Hanson, B. A., & Wong, D. Y. (1984). The harmonic magnitude suppression (HIS) technique for intelligibility enhancement in the presence of interfering speech. Proceedings of the 1984 IEEE International Conference on Acoustics, Speech, and Signal Processing, 18A.5.1-4.

Jex, H. R. (1979). A proposed set of standardized sub-critical tasks for tracking workload calibration. In N. Moray, Mental Workload: Its Theory and Measurement (pp. 179-188). New York: Plenum Press.

Lim, S., & Clements, M. A. (to be published). Windowless analysis of speech for automatic recognition. IEEE Transactions on Acoustics, Speech, and Signal Processing.

Lim, S., & Oppenheim, A. (1978). All-pole modeling of degraded speech. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-26, 197-210.

Lippmann, R. P., Mack, M., & Paul, D. B.(1986). Multi-style training for robust speech recognition under stress. Proceedings of the Acoustical Society of America.

Lippmann, R. P., Martin, E. A., & Paul, D. B. (1987). Multi-style training for robust isolated-word speech recognition. Proceedings of the 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, 705-708.

Mansour, D., & Juang, B. (1988). A family of distortion measures based upon projection operation for robust speech recognition. Proceedings of the 1988 IEEE International Conference on Acoustics, Speech, and Signal Processing, 36-39.

Nakatsui, M., & Suzuki, J. (1970). Method of observation of glottal-source wave using digital inverse filtering in the time domain. Proceedings of the Acoustical Society of America, 47, 664-665.

O'Shaughnessy, D. (1987). Speech Communication: Human and Machine. Addison-Wesley.

Paul, D. B. (1987). A speaker stress-resistant HMM isolated word recognizer. Proceedings of the 1987 IEEE International Conference on Acoustics, Speech, and Signal Processing, 713-716.

Pols, L. (1971). Real-time recognition of spoken words. IEEE Transactions on Computers, C-20, 972-978.

Pols, L., van der Kamp, L., & Plomp, R. (1969). Perceptual and physical space of vowel sounds. Journal of the Acoustical Society of America, 46, 458-467.

Quackenbush, S. R., Barnwell, T. P., & Clements, M. A. (1988). Objective measures of speech quality, Englewood Cliffs, NJ: Prentice-Hall Inc.

Wong, D. Y., Markel, J. D., & Gray, A. H., Jr. (1979). Least squares glottal inverse filtering from the acoustic speech waveform. IEEE Transactions on Acoustics, Speech, and Signal Processing, ASSP-27, No. 4, 50-355.